

格致方法·定量研究系列 吴晓刚 主编

评分加总量表构建导论

[美] 保罗·E.斯佩克特（Paul E. Spector）著
李兰 译 王佳 校

 革新研究理念

 丰富研究工具

 最权威、最前沿的定量研究方法指南

57

格致出版社 上海人民出版社



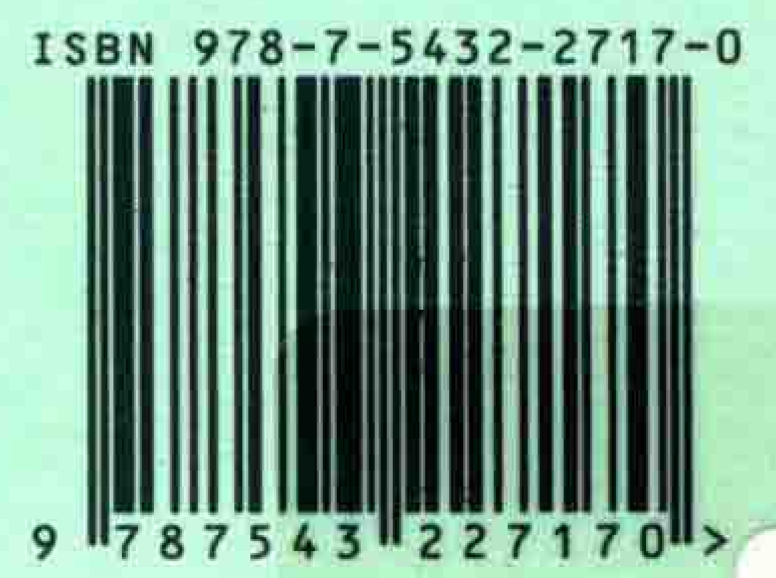
评分加总量表是社会科学研究领域十分常用的工具之一。本书核心论题是如何构建合适的评分加总量表，以正确表达制作者的真正意图，并收到有效的结果。为此，本书不仅考察了多个题项的必要性，还研究了题项答案的合适数量及恰当的题项用语。除此之外，本书还提出了一些题项取舍的原则，如余项系数法、α参数法等，并介绍了量表效度的检验方法和量表信度和标准的处理方法。

主要特点

- 全书结构明晰，简洁扼要，是一本量表设计的入门指南
- 既构建了评分加总量表的设计理论框架，又包含了一些经验之谈，兼具理论性与实操性

上架建议：社会研究方法

ISBN 978-7-5432-2717-0



定价：30.00元

易文网：www.ewen.co
格致网：www.hibooks.cn

您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn


微信


天猫

格致方法·定量研究系列 吴晓刚 主编

评分加总量表构建导论

[美] 保罗·E.斯佩克特(Paul E.Spector) 著
李 兰 译 王 佳 校

SAGE Publications, Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

评分加总量表构建导论/(美)保罗·E.斯佩克特著；
李兰译.—上海：格致出版社：上海人民出版社，
2017.3

（格致方法·定量研究系列）

ISBN 978-7-5432-2717-0

Ⅰ.①评… Ⅱ.①保… ②李… Ⅲ.①社会科学-研
究方法 Ⅳ.①C3

中国版本图书馆CIP数据核字(2017)第017328号

责任编辑 裴乾坤

格致方法·定量研究系列

评分加总量表构建导论

[美]保罗·E.斯佩克特 著
李兰 译 王佳 校

出 版 世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.co)  编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn 发 行 上海世纪出版股份有限公司发行中心	印 刷 上海商务联西印刷有限公司 开 本 920×1168 1/32 印 张 4.75 字 数 80,000 版 次 2017年4月第1版 印 次 2017年4月第1次印刷

ISBN 978-7-5432-2717-0/C·168 定价:30.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于 2011 年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于 2012 年以单行本的形式再次出版发行，共 37 本。我们衷心感谢广大读者的支持和建议。

随着与 SAGE 出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，北京大学教育学院博士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

综观社会科学，研究人员广泛使用评分加总量表。政治科学家为了调查受访者对“政府的信任”，可能会设计好多项问题，然后把各项的得分加总起来，形成一个指标。社会学家为了评估工人的“社会阶层认同”（subjective social class)，可能会问受访工人一组问题，然后把这些问题的答案加总构建一个测量指标。或者像斯佩克特博士(Dr. Spector)那样的心理学家也会在多个同意/反对(agree-disagree)的利克特式测量项（Likert-type items)的基础上，建立一个工作控制点量表（Work Locus of Control Scale)。在评分加总量表应用的每一个范例中，目的都是为了展现个人对某些态度、价值或观点的评估。

构建一个好的评分加总量表并不是一件容易的事情。而且，许多研究生的培养项目并没有专门教授如何构建这种量表。因此，对于那些必须掌握量表构建但尚不熟练的

研究生和那些还需要进修的大学老师，这本专著就具有极大的参考价值。斯佩克特教授清晰且极有耐心地传授我们如何构建评分加总量表的必要步骤。

举个例子：某个年轻的政治科学家正在研究政治价值问题(political values)，比如“自由言论”(free speech)的价值。在调查问卷中，这个学者问受访者是否赞同如下题项：

共产主义者有权利和其他人一样发表言论。

仅仅这一个题项是否足以测量我们关于自由言论的概念？斯佩克特的答案是“不”，然后他便开始了他的解释。在本书中，作者不仅考察了为何多个题项是必需的，还研究了题项答案的合适数量以及恰当的遣词造句等。之后，他提出了一些题项取舍的准则，包括余项系数(item-remainder coefficients)和科隆巴赫的 α(Cronbach's alpha)。完成题项分析(item analysis)后，便可开始对量表的效度(validation of the scale)进行考量。该题项的意思是否就代表了问卷设计者的本来要测量的项目？本书对量表效度的多重标准问题进行了研究，包括来自因子分析(factor analysis)的维度效度(dimensional validity)。下一步是对量表信度(scale reliability)和标准(norms)的处理。

在整本书中，斯佩克特对理论问题的处理相当谨慎。

经验性的检验从来无法证明一项理论概念(a theoretical construct)的成立,但是它们大概能够提供一些证据来支持理论概念的存在。正如斯佩克特教授在本书结论部分所说:“对量表的研究是一项永远不会结束的事业。”

迈克尔 S.刘易斯-贝克

目 录

第1章

绪 论

评分加总量表是社会科学中最经常使用的研究工具之一。它的发明归功于伦西斯·利克特(Rensis Likert, 1932),他利用这种技术来测量态度(attitudes)。这些量表被广泛应用于社会科学研究中,不仅用于测量态度,而且用于测量观点(opinions)、个性(personalities)以及对人们生活和环境的描述。目前量表可用于测量情绪状态(如愤怒、焦虑和抑郁),个人需求(如成就、自主、权力等),个性[如控制点(locus of control)和内向],或者对工作的描述(如角色的模糊性和工作量)。这些只是成百个变量中的一些变量,而量表就是在这些成百个变量的基础上发展出来的。很多变量存在几种量表测量方法,其中的一些量表是为了特别的研究目的而生成的。

一个量表要成为评分加总量表,得具备四个特征。首先,要含有多个题项(multiple items)。加总(summated),顾名思义就是多个题项被合并或者综合。第二,每个单项必须可以测量那种潜在的可定量测量的连续统一体

(underlying, quantitative measurement continuum)。换句话说,它测量的是一种定量而非定性变化的事物特性。比如态度这个东西,它可以从非常赞成变化到极度反对。第三,每项都没有所谓"正确"答案。这个特征使得评分加总量表与考试题中的多项选择题区别开来。因此,评分加总量表不能用于测试知识量或能力。最后,量表的每一项都是一个陈述,要求受访者对每项陈述打分(ratings)。这包括询问受访者在多个答案选项中,有哪一项最符合他们对这个陈述的看法。绝大多数评分加总量表会给出4至7项答案选项。

表1.1是一个关于工作控制点(The Work Locus of Control Scale,以下简称WLCS; Spector, 1988)的评分加总量表例子。WLCS是一个包括16个题项、每个题项包括6个回答选项的同意型量表。关于这个量表,这里有三个注意事项。首先,在表格顶端是包含6种回答选项的数字,从完全不同意到完全同意依序排列。最小值1表示完全不同意(disagree very much)。最大值6表示完全同意(agree very much)。在这些答案下方是问题项或陈述,要求受访者给出相应的回答。在每一题项的右边是所有的6个答案选项。受访者根据自己的情况对应每个题项圈选出一个答案。

WLCS代表了评分加总量表普遍采用的一个形式,但是也存在其他的形式。比如,可以请受访者写下相应的数

表 1.1 工作控制点量表(WLCS)

下面的问题关注你对一般意义上的工作的信仰。这里的工作不是指你现在正在做的工作。

1 = 完全不同意　　4 = 有点同意
2 = 不同意　　5 = 同意
3 = 不太同意　　6 = 完全同意

1. 工作在于你的创造。	1 2 3 4 5 6
2. 对于大部分工作,人们能基本上完成他们开始设定的目标。	1 2 3 4 5 6
3. 如果你想从某份工作中得到什么,那么你就能找到一份能给予你想要的工作。	1 2 3 4 5 6
4. 如果雇员对于他们老板的决策不满意,那么他们应该对此做点什么。	1 2 3 4 5 6
5. 找到一份你想要的工作很大程度上取决于运气。	1 2 3 4 5 6
6. 挣钱主要取决于好运。	1 2 3 4 5 6
7. 如果努力,大部分人能做好自己的工作。	1 2 3 4 5 6
8. 为了找到一份真正的好工作,你需要在高层有亲属或朋友关系。	1 2 3 4 5 6
9. 升职通常取决于好运。	1 2 3 4 5 6
10. 当获得一个真正的好工作时,你认识谁比你知道什么更重要。	1 2 3 4 5 6
11. 那些工作表现好的员工应当获得升职。	1 2 3 4 5 6
12. 如果要赚很多钱,你必须认识对的人。	1 2 3 4 5 6
13. 要在大部分工作中都表现很出色需要靠很多运气。	1 2 3 4 5 6
14. 工作表现很好的人一般都能得到相应的回报。	1 2 3 4 5 6
15. 大部分雇员对他们主管的影响要比他们认为的要大。	1 2 3 4 5 6
16. 人赚钱多少主要取决于运气好坏。	1 2 3 4 5 6

值来表示他们对应于每个题项的答案,而不是在列出的答案选项中划圈。本书接下来的阐述将针对所有的评分加总量表,不论该量表的具体形式。

WLCS是由笔者研究设计出来的一种量表(Spector,1988)。本书将以此为例,因为它正好可以呈现一个量表

设计的各个步骤。当然,我并不因此就认为这个量表在任何一方面要比其他量表更完美、设计更周密、或者其概念效度更好。更确切地说,我只是认为设计该量表所采用的方法是量表设计人员采用的典型形式。

评分加总量表的应用常常是基于以下几个理由。首先,它可以呈现出色的心理测量属性(psychometric properties),即一个好的评分加总量表具有良好的信度(reliability)和效度(validity)。其次,设计一个评分加总量表,相对来说成本较低且较容易。题项的设计比较直截了当,且量表的初步设计大概只需要100到200个受访者(subjects)。最后,对于受访者来说,一个设计优良的量表简单明了,能够很快完成,而且通常不会遭到抱怨。

当然,评分加总量表也有缺陷。也许最大的缺陷在于要求受访者必须具有相对较高的教育程度。如果受访者阅读能力较差,那么他们完成量表一定会有困难。另一个缺陷是,设计一个好的量表得具备一定程度的专业背景和统计技术。然而,我也发现,一些修过一到两门统计课和(或)测量课,并得到一些指导后的本科生,也可以设计出非常好的量表。正如多数事情那样,一旦你知道方法,那么量表设计也就并非难事。

本书的目的就是详尽地阐述如何设计一个评分加总量表。这里写的步骤比较具有代表性,绝大多数研究人员设计量表时都会遵循采用。而多数测试机构(testing

firms)，比如 ETS(Educational Testing Service)和 PC(Psychological Corporation)所设计的测试项目，常常涵盖并利用大得多的受访样本。但是，它们的基本方法也和本书所述非常相似。

本书将尽可能详尽地涵盖量表构建的所有必要步骤，并且会提出一些建议，以便读者在设计量表的过程中少走弯路。对于第一次接触量表设计的研究人员来说，仅仅学完本书，并不足以指导完成量表构建。我推荐读者再向那些有经验的量表设计者请教。至少，应该找个人能帮你检查相关的步骤、题项和得到的分析结果。

第1节 | 为什么采用多项量表?

评分加总量表的设计需要投入很多的时间和精力。它也需要受访者能够花上点时间来完成他们的评价。因此,一个问题就自然而然出现了:为什么要搞得这么麻烦?为了知悉某人的观点,为何不直截了当地向他/她本人提出“是/否”这样简单直接的问题呢?

为何“是或否”的简单问法并不合适?这里涉及三个合理的理由:信度、精确性(precision)和范围(scope)。单个“是或否”的问题项,不能得到人们稳定的反馈。一个受访者也许今天回答“是”,而明天回答“否”。因此单个“是否”的问题项特别不可靠。而且也不够精确,因为它将测量仅仅限制在是和否这两个层面上,导致受访者只能被分为两个组,而没办法在每个组内进一步区分。最后,许多待测量的特性变化的范围都很广,无法利用某个“是/否”的问题得到确认。有些问题很复杂,要评价测量它们往往需要多个题项。

这些问题最好用个例子来说明。一个经常研究到的

话题是民众对政府的态度。为了评价这种感觉，我们可以提出一个简单的问题，如：

您喜欢这个政府吗？（是/否）

很不幸的是，所有回答“是”的受访者的感受程度是不一样的。一些人也许谈得上热爱这个政府，而另一些人则只是有那么一点点喜欢。同样地，有些回答“否”的人是憎恨这个政府，然而另外一些人也许只是有那么一点点反感而已。那些感觉模糊的左右摇摆人士，就不得不去选择一个“是”或“否”，从而被归入了那些具有强烈感觉的人群之中。因此，这样的问话形式，对于大多数研究目的来说就不够精确了。

有多种情况会导致人们的回答随着时间的推移变得不可靠或者不一致。首先，那些感觉模棱两可的人可能根本上是随机地回答这个问题。他们可能会根据日子、个人情绪和天气回答“是”或“否”。假如某个感觉模棱两可的人在不同的场合，被问到同样的问题，那么我们将会观察到他/她的不一致性。实际上，在心理物理学(psychophysical)意义上，回答“是”和“否”各占50%就意味着感觉模棱两可了。[1]

受访者错误回答也会导致不可靠。他们也许想说“是”但结果说成“否”，他们也许误读了题目(比如，我不喜

欢政府)，或者误解了问题本身。他们也许对问题本身的意思把握不定。“政府”是指联邦政府么？还是州政府，抑或是本地市政府？或者涵盖这三种政府？所有这些因素都会引起误差，从而导致不可靠。

最后，还有一个困难之处还在于人们的感觉并非如此简单明了。他们也许喜欢政府的某些方面，但是不喜欢其他一些方面。特别是当这个问题涉及各级政府时，人们很难回答。更为重要的是，单个问题会过度简化人们的感觉。

这里，评分加总量表的两个特征能够解决这些问题。首先，利用两个以上的选择项来提升精确性。比如问：

请问您感觉政府怎么样？

然后列出以下几个选择项：

热爱

喜欢

无所谓喜欢不喜欢

不喜欢

憎恨

这样的话，那些怀有强烈个人情感的人群和那些较为温和的人群就能得到区分。而那些感觉模棱两可的人也能够

做出自己的倾向反应，不至于被那些态度确定的人群所掩盖。精确性就大大提高了；而且假如选择项更多的话，精确性还能进一步提升。当然，如果受访者能够清晰地回答“是”或“否”，那么多余的选项也不起什么作用。这里的关键是不要超出受访者的能力，给出过多的选项。

多个题项能解决这三个问题。人们会对关于不同层面政府的题项做出回应。问题可以是，他们是否喜欢总统、国会、最高法院和公务员？或者可以问他们关于公共服务、税收、钱是如何花的、政府运转得如何等各种问题。这些问题延展了所要测量概念的范围。该范围可以因问题的选择而扩大或缩小。

多个题项会将测量的随机误差平均化，从而提高信度。假设有 20 个题项，如果某一受访者仅在一个题项上犯错，比如将“热爱”错划成“憎恨”，那么这对总得分（所有题项的总和）的影响是微乎其微的。实际上，在一个方向上的误差倾向于抵消在另一个方向上的误差，从而得出一个随时间变化相对稳定的总分。关于信度，我在后面还会有更为详细地阐述。

最后，多个题项的精确性会更高。对于一个只有 5 个回答选项的问题，受访者能根据他们的答案被分成五组。如果设定 20 个有 5 个回答选项的问题，那么便有 61 种可能的得分（取值范围从 20 到 80），或者说，精确性扩大了 16 倍（至少理论上如此）。

第2节 | 什么是好的量表?

一个好的评分加总量表是既可信(reliable)又有效(valid)。信度可以从两方面来考虑。首先,重测信度(test-retest reliability)意味着一个量表随着时间变化仍能得到一致的测量结果。假定关注的概念(construct of interest)未变,那么每个受访者在重复测量中应该得到大致相同的分值。第二,内在一致性信度(internal-consistency reliability)指用来测量同一概念的多个题项存在相关关系。一个量表也许只能用来说明其中的一种信度。本书第5章和第7章将会详细阐述这两种信度。

信度确保一个量表能一致地测量事物,但它并不保证所测事物是该量表设计所要测量的概念。这个特性(即量表测量其所要测的概念)是效度(validity)。效度有多种类型,有几种基本的方法可以用来建立效度。第6章将探讨这些内容。对一个量表来说,信度和效度是其根本的要素。关于这些,读者可以参阅Carmines和Zeller(1979)的著作。

同时，一个好的量表还需要注意其他事项。首先，题项应当书写清晰、意思确定单一。许多量表出现问题，都是因为题项模糊矛盾，或涵盖多重意思。除非绝对需要，应当避免使用行话。就调查人群和时间而言，口语性的表述也会限制量表的使用。

还有另外一个要素不容忽略，那就是一个好的量表应该适应使用群体的总体水平。举例来说，设计量表时必须考虑群体的阅读水平。为了让一个量表具有广泛的适用性，量表用语要简短精练、意思直接明了。最好的题项应该是那些采用具体而非抽象的概念表述、不让受访者反复猜测琢磨原意的题项。不要让受访者因为不理解某个用词而误解了题项本身的意思。第 4 章将会讨论如何撰写一个好的题项。

最后，设计一个好的量表会考虑到各种可能的偏差因素。私人性质的敏感话题可能会引起某些受访者的自我防卫意识。测量个人心理调节和精神病理学方面的量表非常容易遭到某些受访者心理上的防备，从而被曲解，这点已经众所周知。第 4 章将会阐述关于测量偏差的问题。

第3节 | 量表构建的步骤

评分加总量表的构建是个多步骤过程。一个完整的过程将包含多项独立的研究。图1.1展现了这个过程的五个主要步骤。首先,在构建量表前,必须清晰且准确定义它的目标概念(construct of interest)。如果一个研究者要用量表来测量什么都没有搞清楚,那么不可能设计好一个量表。这看起来是个相当简单的要求,但正是在这个步骤

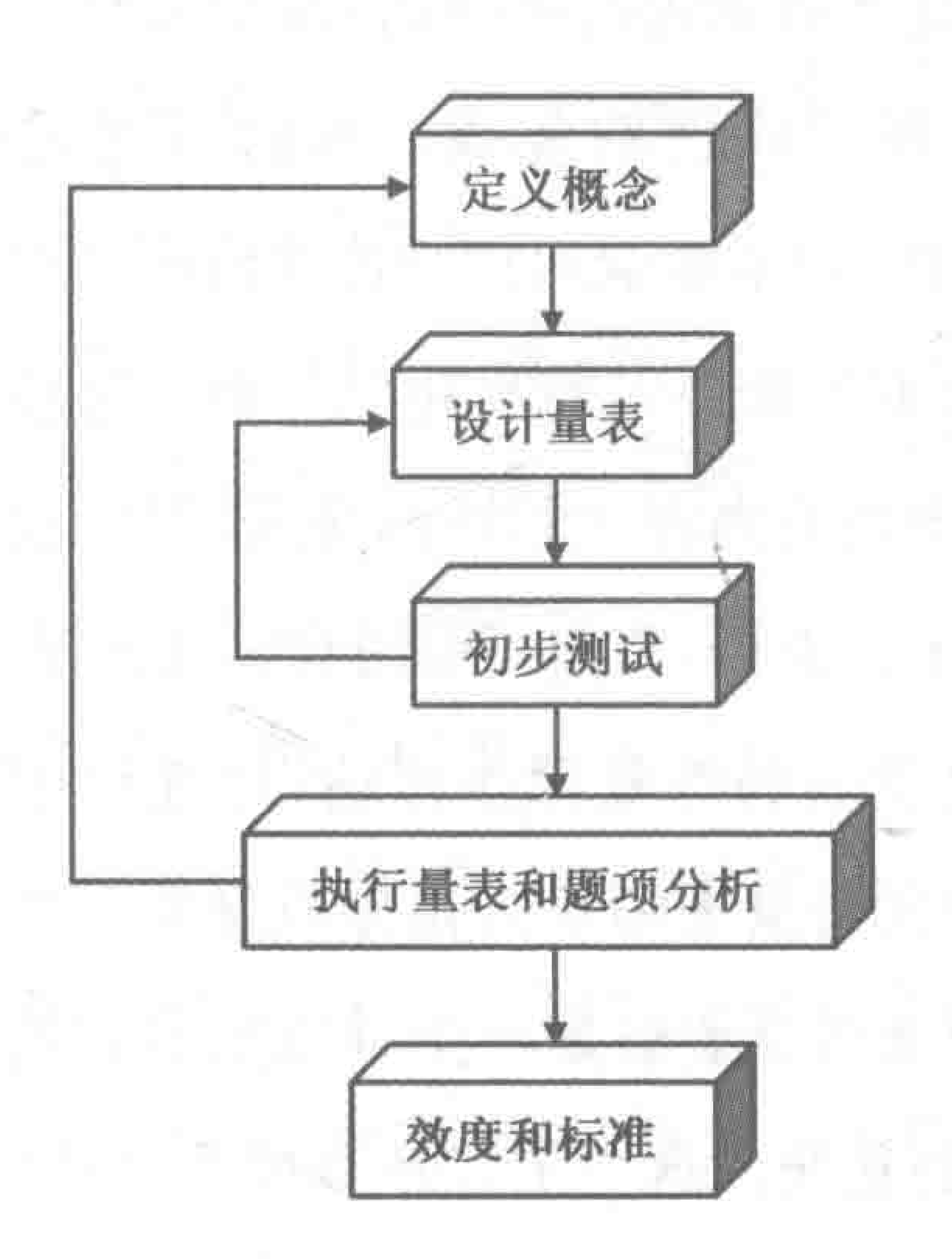


图1.1 评分加总量表设计的主要步骤

上许多量表设计都偏离了方向。太多的量表设计者在定义和完善目标概念上根本没有花费足够的心思。

第二便是量表本身的设计。这包括决定量表的形式、应答项的选择以及量表填写指南的书写。题干的书写也包含在这个步骤。这里的想法是先写一个初步的题项备选库。这个备选库将用于后面的统计分析步骤。

第三，量表初稿要交给一小群受访者进行测试，让他们提出批评意见。他们应当指出哪些题项是含糊的或令人费解的，哪些是不能根据所选择的维度来排序的。这个量表初稿应该根据这些受访者的反馈进行修改。

第四，要对量表进行第一次全面测试(full administration)和题项分析。要选择 100 到 200 个受访者来填写量表。然后对这些数据进行项分析，从而选择一组题项组成具有一个内在一致性的量表。我们可通过计算系数 α (Cronbach, 1951)来衡量内在一致性信度。在这个初步阶段，信度的基本特征得以初步建立。如果这些题项能够成功地产生一个具有内在一致性的量表，那么我们就可以继续最后一步了。否则，我们还得回头去修改量表。

第五，对量表的有效性作评估，并进行标准化。传统上，效度被定义为量表测量其目标概念的一个特性。换句话说，一个有效的量表就是一个可以测量它设计之初想要测量的概念的量表。这个定义稍显简单，但这里如此定义也没问题。后面我们还会继续这个话题。

在这一步，我们应该进行一系列的效度研究，以确认量表能如预期进行。这一步特别像理论检验（theory-testing），因为假定了量表与其他变量之间的关系。然后，收集数据用于验证理论假设。当支持效度的证据收集充足后，我们便有信心认定该量表能够测量预期中的理论概念。

在收集效度数据的同时，标准化数据（normative data）也被收集了。标准（norms）描述给定人群应用一个量表时的分布特征（distributional characteristics）。个人量表的得分可通过其与总体分值的分布关系来解释。大规模受访者样本可以用来估计总体的分布特征[比如均值和标准差（mean and standard deviation）]。

上述五个步骤是构建量表的基本过程。但不幸的是，许多量表设计者在第一步和（或）第五步下的功夫不足。毫无疑问，这是因为这两步是最难的。它们都仰赖扎实的概念和理论思考，以及相关的文献研究。效度包括对量表的理论假设进行测试等一些研究工作。一个好的量表设计者必须首先是一个好的研究者。

本书剩余部分将从量表的基础——经典检验理论（test theory）开始讨论，阐述设计一个评分加总量表的所有步骤。图 1.1 所示的每一个步骤，包括定义概念、设计量表、初步测试、进行题项分析、验证量表以及建立标准和信度，我们都会一一讲到。

第2章

评分加总量表的理论

在设计评分加总量表之前，有必要简要地回顾一下其背后的一些理论。隐含其中的基本理念来自于经典检验理论(classical test theory)，它为可重复的加总测量提供了理论依据。

经典检验理论将真实值(true score)区别于观测值(observed score)。真实值是指每个受访者在被测概念或变量(construct or variable of interest)上具有的理论值。观测值则是通过测量实际所获得的值。现在假定每个受访者在所测概念上都有一个真实值，但是这些真实值是不可能被直接观察到的。相反，它们是从观测值中推断出来的。如果某人有非常可信且有效的测量，那么观测值就等同于真实值。

根据经典检验理论，每一观测值由两部分组成：真实值和随机误差(random error)，即：

$$O = T + E$$

其中，O 是观测值，T 是真实值，E 是随机误差。随机

性的误差是假定来自均值为零的总体的。这意味着随着测量次数的增加，误差将趋向于平均化为零。

在评分加总量表中，每个题项都被设计成某个待测特征(intended trait)的观测。每一题项都代表着对一个真实值的独立评估。如果对所有个体题项进行加总或平均，那么测量误差也就可以被认为平均化为接近于零值，得到一个接近真实值的估计值。

测量误差与信度呈负相关关系。在任何给定的测量情形中，误差成分越大，信度也就越低。此处以一单个题项测量某一特性为例。利用单一题项进行测量是特别不可靠的，因为它们包含很大的误差成分。如果误差是随机的，那么有时它们会扩大或缩小真实值的观测值。当测量重复进行时，观测值将出现不一致(不可信)[inconsistency (unreliability)]。当把多个题项加入到对某一真实值的估计时，误差将会被平均化并趋向零，从而得出一个更加准确且前后一致(可信)的测量值。

因此，提高信度的一个方法就是增加题项。这正是评分加总量表背后的基本原理——利用足够多的题项来获得一个合理水准的信度。单个题项的误差越多，一个总体量表就需要设计越多的题项以获得好的信度。如果题项足够的话，要获得一个总体可信的量表，单个的题项并不要求非常可信。当然，研究者应尽量构建一个高质量的题项。如果仅仅依赖一大堆质量低劣的题项，就以为能够将

误差平均化至零，这种想法是错误的。虽然这样可能能够获得想要的信度，但是低劣的题项也许不能被认为是有效的。

信度的获得仅仅意味着单个题项的误差成分已经被平均掉。但是，多题项的应用并不能确保所测的真实值便是所要的真实值。有可能（在许多情形下是非常有可能的）所测得的真实值并不是量表设计要测的特点。

经典检验理论是评分加总量表以及其他多种测量背后的基本原理。然而，经典检验理论往往过于简化，并没有将一些会影响人们对量表的回答的因素考虑在内。经典检验理论的基本公式可以加入一个额外成分：

$$O=T+E+B$$

其中，B 表示偏差，它由那些导致观测值偏离真实值的系统性影响因素构成。系统性影响因素（sytematic influence）不是随机的，不来自均值为零的分布。因此，它们不会因为多个题项而被平均掉。偏差代表一个或多个影响观测值测量的替代性特征（alternative trait or traits）。

一个最为麻烦的偏差源是社会期望（social desirability）（Crowne and Marlowe，1964）。社会期望（以下简称为 SD）是一种倾向，指某些受访者在回答题项时，是根据社会普遍期待的或可接受的方向来回答，而不是给出自己真实的感觉或者答案。比如，SD 倾向强的人们比较不太可能承认

他们自己曾经“玩纸牌时作弊”或者“偷窃家人的东西”。因此，对于某些人来说，观测值也许只能反映SD的值，而不反映感兴趣的特征；或者也许两者都有反映。

在个人心理调整和精神病理学的测量中，这个问题尤其凸显。许多测量这些概念的量表都含有一些不恰当的题项。那些关于奇特想法或行为的题项（比如“你听到了声音吗？”或“你享受将痛苦强加于别人的行为么？”），就有可能存在问题。如果人们在这些量表中得分很低，要么是因为他们的这些特征的真实值本来就很低，要么是因为他们的SD（社会期望）倾向性强，从而不愿意承认这些社会普遍不认可的事情。

已经有相关研究关注量表偏差的几个来源。其中有被称为反应定式（response sets）的，它指受访者对题项系统性的反应倾向。举例来说，默认反应定式（acquiescence response set）指某些人不论内容如何一概对题项表示同意的倾向。默认反应定式倾向强的人将会给所有题项高分。因此他们的给分是有“水分”的（假设题项都是正面措辞的）。而且这种“水分”是很难变化的，不可能仅仅通过简单增加题项来进行处理。

现在已经有些方法可以用来处理一些可知的偏差来源。尽管偏差不可能完全消除，但是它们可以被减少。比如，对于社会期望，可以通过仔细斟酌、恰当使用题项的字句，减少其中社会期望的内容，从而降低这一偏差源的影

响。另外一些量表形式，比如强行选择(forced choice)，也可以用来处理社会期望问题。

不幸的是，我们目前不大可能知晓所有的偏差来源。我们永远不能确定的是：这些系统性的影响因素并没有影响到我们的测量。量表设计者一直在假设：经典检验理论代表了一种合理的、非常接近他们测试条件的情况。然而，他们也承认他们的量表有可能受到偏差的影响。因此，如果要证明量表测量的是原定目标，而不是偏差，那么效度起到非常关键的作用。

第3章

定义概念

量表设计最为关键的步骤之一就是定义概念。此处无需赘言，如果一个概念的性质未被清楚描述，我们不可能设计一个量表来测量它。当一个量表出问题时，原因经常在于设计者没有认识到认真和明确描述目标概念的重要性。没有一个清晰定位的概念，就很难写好题项，也很难为验证效度而提出某些假设。

社会科学研究的一个难处是许多概念是理论上的抽象，没有可知的实体现实。这些理论抽象也许处于不可观察的认知状态，无论是个体的（比如态度），还是共同的（比如文化价值）。社会科学家也许清楚这些概念，但是受访者就不一定搞得清楚这些概念了，无论这些受访者是个体还是庞大的社会实体。

假如一个概念是理论上的抽象，那么我们怎么能够确定某种量表可以测量这个概念呢？这是个很难解决的问题，毫无疑问阻碍了社会科学的发展。效度验证是可能的，但是它必须在评估一个概念有用性的清晰环境下才能

进行，而且该概念与其他某些可能更具客观观察性的概念之间存在可能的理论关联。一个概念不可能完全独立，它是一个更宽广的理论网络的一部分，这个理论网络描述许多概念之间的关系。一个概念不可能在真空中得到发展。关于效度验证的问题，我们将在第 6 章进一步探讨。

在多数情形下，量表设计者对概念的构建花费的精力并不够，也许是因为他们自以为对所需构建的概念已经有一个很好的主观认识了。这样的想法是非常危险的。如果构建的概念事先没有得到清晰定义，那么量表的信度和效度就非常有可能很差。也就是说，构建的概念与量表之间的关联就不清楚。

这里强烈推荐一种归纳式(inductive)的方法。量表设计从一个清晰界定的概念开始，通过对这个概念的界定指引随后量表设计的过程。大部分设计工作，尤其是效度验证，采用确定法(confirmatory approach)，即在理论观念的指引下进行效度确认。然后得到关于量表与其他变量之间的关系的理论假设。效度研究将检验这些假设。

有些量表设计者采用演绎式(deductive)的方法。题项被发放给受访者，然后设计者运用复杂的统计方法(比如因子分析)来试图发现这些题项中的概念。这其实是一种探索性的方法，其中有关概念的工作主要集中在解释结果，而不是形成一个先验的假设。使用这种方法必须特别小心。几乎所有相关的题项都必然会得出一些能赋予意

义的因子。这里的问题是,从题项分析阐释得来的概念也许太过明显,而非真实概念。

一个同事曾经给我讲过一个故事来说明研究人员采取探索性方法时必须非常谨慎。当他还是个研究生时,他见几个相当资深、富有经验的研究人员复查一个因子分析的结果。而后他们对于得到的结果解释非常满意,互相庆祝,因为那时这个结果似乎非常有意义。这时,那个负责分析数据的研究助理进来了。他非常尴尬,说是打印出错了。他们看的那些数据不过是些随机数。[2]对于探索性结果的解释必须得万分小心。

这并不是因为因子分析或者探索性研究中有无法避免的内在错误。然而,对于一项测试的构建来说,我们还是推荐归纳式的方法。效度验证是个非常微妙且困难的事情。当一开始未能建立一个扎实牢靠的概念基础时,这种验证就显得尤其艰难。本书第 6 章将会谈到,在量表设计中采用归纳法时,因子分析作为一个效度验证的策略能非常有用。

第1节 | 如何定义概念

定义概念也许是量表构建中最为困难的部分。当要构建的概念特别抽象和复杂时,这项任务尤为困难。概念的构建工作应当从一个概念的一般定义开始,然后逐渐细化。一个概念描述得越是清晰,那么我们设计测量它的题项时就越容易。

在描述一个概念时,借鉴已有的概念和量表设计工作是非常有帮助的。除非要构建的概念是全新的,否则在已有的文献中肯定有所讨论,可能还有些经验性研究。并且,也许还有现成的量表可用于之后的评估。现有的文献应该成为定义概念的一个工作起点。先前的概念界定和可操作化能够为自己的工作提供一个坚实的基础。一项量表设计工作常常为先前文献中某个尚未得到很好研究的常见概念提供进一步的改善和提升。

概念定义的第一步是文献综述。我们应当首先仔细阅读有关概念的文献,特别注意描述那些概念的细节。如果一个概念非常常见,那么它有多种不同的定义是很有可

能的。为了设计一个量表,我们必须选择一种定义。毫无疑问,一个概念的不同定义都会在一个更为宽泛的理论背景中得到讨论。一个概念不可能在真空中得以阐释,它得存在于与其他概念之间的关系网络中。如果概念/理论工作做好了的话,不仅仅题项写起来轻松,而且效度验证的框架建立也会有的放矢。

此处以"压力"(stress)这个概念为例。有许多关于压力的理论,从而得到关于什么是压力的多个不同概念定义,以及许多压力量表。一些研究人员认为压力代表特定的环境条件。家庭成员的死亡,或者繁重的工作,都代表压力。另外一些研究人员则将人们情绪上的反应视为压力。感觉沮丧(也许是由于家人的死亡,或者工作繁重)便是压力。还有些人将生理上(physiological)的反应视为压力。根据这种定义,心率增加、血压增高、免疫系统受到抑制全都属于压力。因此,用来测量压力的确切程序或量表将依赖于该压力的准确界定。环境可以通过外部观测来测量,情绪可以通过个人的自我报告来测量,而生理上的反应可以通过适宜的医学仪器来测量。无差别的一般性方法无法用来测量上述三种概念,因为它们虽然都被称为压力,但其实代表不同的含义。

试图准确定义压力陷入了困境。开始,研究人员试图采用一个单纯的环境定义,并且厘清什么样的环境条件可以被称为压力,而什么样的不能被称为压力。家庭成员的

去世可以认为是一种环境压力。为了处理对家人死亡这样的事件的个体差异，一些研究人员就扩大他们对压力的定义，将生者对逝者的情感也纳入到定义中。如果生者憎恨这个家人且对他/她的死感到高兴，那么他/她的死亡就不能认为是压力的一个例子。如果要认定家人死亡是压力，那么生者就应当对其感到难过。如果生者的感觉很重要，那么个人的情绪反应就应该成为压力定义的一部分，从而该定义就同时涵盖了环境和情绪。现在定义就产生变化了，不单纯是指环境，还包含了个人对环境的反应。

诚如上述所见，概念定义的过程很复杂，甚至变得非常麻烦。这里特地选择压力作为例子来讲，是因为它比较麻烦，足以挑战所谓我们想象中完美的概念开发。正因如此，许多压力研究者直接放弃把压力作为一个概念。但是，压力常作为一个主题领域的名称来使用。研究策略集中于确定多个具体概念间的关系。家庭成员的死亡、对逝者的感受、对死亡的反应都能够定义为不同的概念，而这些不同的概念可以据其与其他感兴趣的变量之间的关系得到研究。比如，研究集中在家庭成员的死亡是如何影响在世者的健康。

如果用于测量感兴趣的概念的量表已经存在，那么这些已有量表的内容就对量表的设计有帮助。根据现有的量表设计出新的量表，是很常见的。在一些尚无高质量量表的领域，我们常常看到这种情形。现有几个量表的题项

可以作为新量表题项备选库的设计起点。这些题项将被修正，而且更多新的题项也添加到这个题项备选库中。然后，根据这个题项备选库得到一个最终的量表。

最为困难的情形也许是，关于一个概念，没有任何现成的理论或经验性成果可以参考。由于没有任何基础，概念的形成和量表的设计可能同时进行。在量表设计的过程中可能要经过多次反复尝试，概念才能构建得完善，能用于量表设计。

第2节 | 概念的同质性与维度

概念的定义能从非常具体且狭义的变成多维度的。某些概念非常简单，它们的内容用一条题项就足以涵盖。另一些概念就很复杂了，因此它们可以分成多个子概念(subconstructs)。复杂概念的内容也就只能依靠含有多个子量表(subscales)的量表来呈现了。

个人对消费产品的感觉是个非常同质性(homogeneous)的概念。我们可以邀请一些人来品尝一种新的薄饼，然后问他们是否喜欢这种产品。对一个市场研究者来说，如果喜欢就意味着将来愿意购买，那么这个概念的细化程度(level of specificity)就已经足够了。但是如果要做别的用处，这样的细化程度也许就不一定够了。喜欢也许可以继续细分，比如喜欢口味，喜欢质地，喜欢形状，喜欢颜色，以及喜欢气味等。这么一个简单的概念都可以细化。

其他许多概念都要复杂得多。有研究表明工作满意度(job satisfaction)由多个成分组成，且其中的大部分成分之间的相互关系不大。雇员可能对工作的某些方面满意，

但对其他方面不满意。一个人也许对其收入满意，但不喜欢他/她老板。另外一个人也许喜欢工作本身，但不喜欢他/她同事。大部分测量工作满意度的量表都由子量表来估量这些组成要素，虽然不同量表设计者选择不同要素。

概念定义工作的一部分就是决定一个概念如何细分。在多题项量表中，我们可以将每一个题项都当成是概念的一个独立维度或者方面。然而，评分加总量表的总体理念应该是各题项综合在一起分析，而不是独立分析。甚至在设计多题项子量表领域，关于一个概念应该分成多少个不同方面来进行定义，不同的量表设计者往往很难达成一致意见。

关于应当如何细化概念这一问题的终极答案，必须建立在理论和经验效用的基础上。如果对一个概念进行细分能够明显增加一个理论的解释力，并且如果实证研究上能够支持，那么就可以进行细分。如果理论问题因此变得太过复杂笨拙，或者实证上的支持也不足，那么还是不要细分为好。在科学上，简约原则(principle of parsimony)应该受到遵循，即在同样水平的解释中，应当采纳那个最为简单的解释。

第3节 | 工作控制点的理论发展

在心理学和其他社会科学领域，控制点（locus of control）是个非常流行的个性变量（personality variable）。罗特（Rotter, 1966）将控制点定义为对生活中强化（reinforcement in life）的一个普遍预期。一些人相信强化（奖励和惩罚）被置于他们的自我控制之下，一些人则没有这种信心。虽然控制点被作为一个连续体进行测量，但是理论上，内控型（internals）与外控型（externals）是有区别的。内控型的人相信他们具备个人的控制力，而外控型的人则认为运气、命运或者其他有权力的人在掌握着他们的强化（reinforcement）。

遵照法里斯（Phares, 1976）的建议，我决心设计一种量表来测量工作环境下一个特定领域的控制点量表（a domain-specific locus of control scale）。首先的工作就是回顾有关控制点的文献资料，重点关注对工作环境的有关研究。我对文献中内控型和外控型的人的特征和行为进行了细致的考虑。有些特征完全是从理论上分析出来的，而

有些是经过比较两种个性类型的研究得来的。我也对与工作领域有关联的特殊类型的强化因素(reinforcers)进行过研究。

依照工作的界定,工作控制点这一定义从一个更为宽泛的概念延伸出来,重点关注工作中有关强化或者激励控制的普遍期待(generalized expectancies)。内控型的人感觉他们能够控制工作中的强化因素,而外控型的人则感觉他们做不到。外控型的人将控制他们的因素归为运气、命运,或者其他有权力的人,大多都是指上司。根据概念的有关描述,及内控型和外控型人格特征的具体表现,我设计了有关题项。与压力测量量表这一例子相比较,这个课题并不很难。这里的优势在于,作为工作控制点量表设计基础的一般性概念本身已经发展得很好。有丰富的理论和大量相关文献可供查阅。但是,这并不能保证设计出来的量表就是科学的或者实用的,或者就能够测定预期目标的。概念化(conceptualization)的价值只有通过效度验证以及随后的量表应用研究,方能显现出来。

第4章

量表设计

概念界定如果完成的好，就能顺利进入量表设计的下一步。这里有三个部分的工作需要完成。首先是答案选项的内容和数量，其次是题项本身，最后是需要专门给受访者的操作指南。

第1节 | 答案选项

构建答案选项时首先需要决定的是受访者要回答的内容。三种最为常见的方式是：同意(agreement)、评价(evaluation)和频率(frequency)。同意便是询问受访者指出他们对题项的同意程度。评价是请求受访者对每一题项的内容给出一个相应的评价分数。而频率则是请求受访者对每一题项所述事项具有、应有或者将发生的频次给出一个判断。

同意型答案选项通常是围绕一个中性点的两极对称的，要求受访者对每一题项给出同意或不同意的回答，以及他们同意或不同意的程度(magnitude)。答案选项也许会要求受访者在"强烈""一般""有点"同意和不同意中做出选择。这些修饰语(modifiers)在同意和不同意选项中相同，使得答案选项呈现对称。许多量表设计者还会增加一个中间选择，诸如"既不是同意也不是不同意"，虽然这样的选项并不是必需的。

斯佩克特(Spector, 1976)曾计算有关同意、评价和频

率的流行修饰语的心理量表值。他让大学生对有关修饰语给出排序，然后采用吉尔福德(Guilford, 1954)的数学模型将排序数据转化为心理量表值。表 4.1 表明，这三种答案选项都具有大致均匀分布的同意修饰语。虽然修饰语均匀分布并非必要(Spector, 1980)，但这样做可能让受访者填表更轻松。

表 4.1　同意、频率和评估量表的答案选项

同意		频率		评估	
答案选项	量表值	答案选项	量表值	答案选项	量表值
有点	2.5	非常少	1.7	非常差的	1.6
一般	5.4				
倾向于	5.4	很少	3.4	差的	3.6
非常	9.1	有时	5.3		
		偶尔	5.3	过得去的	5.5
		经常	8.3	好的	7.5
				非常好的	9.6

同意答案型选项特别灵活，是最为常见的。此类型的题项可以用来测量不同类型的变量，包括态度、个性、观点或者对环境的报告。表 4.1 提供 3 个选项，这将产生一个 6 点量表(six-point scale)。

评价式选项请求受访者依照由好至差的方向对题项给出评分。表 4.1 的选项由肯定(非常好)延伸至否定(非常差)，并且没有中间选项。评价答案选项的形式可用于测量态度，或者评估行为表现。例如教员评价表，它常请求学生从多个方面对老师做出评价。

频次量表要求受访者回答某事发生或应当发生的频率或者次数。一些研究人员认可给出数值选项(numeric anchors)——比如每天一次或每天两次——这种形式的优越性(参见 Newstead and Collis, 1987),但是大多数量表似乎还是用文字选项(verbal anchors)。表 4.1 所含频率选项从极少(rarely)到经常(most of the time)。一些量表采用从不(never)和总是(always)来界定量表的两个极端选项。频率答案选项一般用于测量量表中的个性特征,这些量表让受访者标明他们做出某些行为的频次。这类选项也被用于测量环境特点,让受访者指出某些事件发生的频次。

对许多的概念来说,这些答案选项的任何一种都可用得上。但对于其他一些概念来说,某种答案选项也许要优于其他的。假设某研究者对人们的选举行为感兴趣。为了确定人们参与某些与投票行为有关的行为的频率,通常情况下采用频次题项可能是最为合理的。例如,该题项可能会问:

您经常参与大选时的投票吗?

答案选项可能是:“总是”“有时”或“从不”。

这个问题也可被处理成同意选项型,但如此处理并不合适有效。试想如下的题项:

我总是参加大选投票。

我偶尔参加大选投票。

我从不参加大选投票。

受访者将填写对每一题项的同意程度。只要他们的答案与其他形式项下的答案一致，只选择同意其中的一项，那么通过这种形式所获得的信息与其他形式也是相同的。但是，同意型选项形式要求更多的题项，并且如果受访者同意（或不同意）某些看起来互相排斥的问题项时，就会产生矛盾模糊的结果。

评价答案形式也可以被采用，但是同样效果不好。比如可以这样问：

请问您在大选时的投票记录怎么样？

这里的问题是如何去衡量受访者投票记录的好坏。有人也许认为参与一半的投票就是个好的记录了，但另外一个人也许认为这样的记录是坏的。

频次答案形式在一定程度上也存在同样的问题，因为并非所有人对答案选项的理解完全相同（Newstead and Collis, 1987）。但是，频次形式可以更为直接获得人们对于问题当中行为发生的次数，而评价形式则可以获得人们对于某种行为的感受。目标概念本身的内涵将决定采用

哪种形式更为合理。

另外一个需要确定的事情是答案选项的数量。大家也许会假定答案选项越多越好，因为选项越多，结果可能越精确。这当然没错，而且确实有量表设计者曾经采用过的答案选项超过 100 个。但是我们必须考虑填写量表的人的测量敏感度（measurement sensitivity）。当答案选项增加，效益递减点（point of diminishing returns）会很快到来。虽然不同的观点会有一些小小的差异，但一般来说 5 到 9 个选项对于大多数量表来说是个较优选择（如，Ebel，1969；Nunnally，1978）。

表 4.1 有助于选择那些大致均等分布的答案选项。对于每个答案选项，该表也提供了其量表值。但该表并不包含所有可能的选择类型，或者是必要的最优选择。它们只是作为类型选择的一个起点，在这里提出来。更多的模式请参见斯佩克特（Spector，1976）。

第 2 节 | 量化答案选项

我们选择答案选项，以便它们能够依照一个测量连续统一体（measurement continuum）进行排列。频次一般依照由从不发生（nonoccurrence）（无，或从未发生）到经常发生（constant occurrence）（总是，或连续发生）的顺序变化。评价则依尽可能的差（as poor as possible）到尽可能的好（as good as possible）的顺序变化。同意则设定为两极型的（bipolar）答案，从完全不同意变化到完全同意。无论这些答案选项是什么，它们都必须依照由低到高的顺序进行排列，并且每一选项都标上数值。

对于有些概念，它们的赋值也许会从零到一个高的正值进行变化。这类量表是单极型的（unipolar）。另外一些概念，它们可能既有正值，也有负值，零值大概处于中间位置。这类量表是两极型的（bipolar）。事件发生的频次是单极型的，因为世界上各种现象发生的次数不可能低于 0 次。态度常常是两极型的，因为一个人可以有肯定、中立或者否定的态度。

对于单极量表来说，答案选项是按照由低到高（从1开始）的顺序进行连续编码。因此，一个5点量表（five-point scale）会从1编到5。两极量表也可按同样方式编码。一些两极量表同时采用正负值来进行编码。一个6点量表可以从－3编到＋3，用负值来表示不同意的答案，用正值来表示同意的答案。如果有中立答案，那么该答案可以标为0。

一个量表的总分值，可以通过将各个题项的答案相加而得以计算。如果既有正面措辞项，也有负面措辞项，那么负面措辞项的分值必须作逆向计算（reverse scored）。否则，这两种类型的题项将会互相抵消。对于一个赋值由1到5的量表来说，负面措辞项的分值将会被逆转。因此，5＝1，4＝2，3＝3，2＝4，1＝5。

下面的公式可以完成这种逆转：

$$R=(H+L)-I$$

其中 H 代表最高值，L 代表最低值，I 是某题项的答案，R 是计算出来的逆转值。以一个五点量表为例，如果某项陈述得到的答案值为2，那么

$$R=(5+1)-2$$

或

$$R=4$$

第 3 节 | 题项主干的编写

量表设计的第二步便是编写题项主干。题项主干的叙述，相当程度上依赖于要求受访者做出的判断或反应的类型。同意型题项是个陈述性的表述，即某人可对其表示同意与否。比如下面的例子：

死刑应当废止。

我喜欢听古典音乐。

我在陌生人身边感到不舒服。

频次型题项则经常与事件、环境或者行为有关，意味着它们发生的频率。受访者也许会被问到下面这些情形发生的频率：

总统候选人会做出一些他们明知不能兑现的竞选诺言。

你的运动剧烈到足以提高你的心率。

你的丈夫帮忙干家务。

评价型题项则常常使用一些表示人物、地点、事物、事件或行为的词语或短语，从而人们可对它们进行评价。下面的例子大概就是待评价的题项：

您所在社区的警察服务。

您刚试用的面巾纸的柔软度。

您最爱的运动团队上周表现如何。

一个好的题项应当清晰、简洁、不模棱两可，并且尽可能具体。它也应当与答案选项的性质相配。编写好的题项是量表设计的一个核心部分。下面是题项编写中应该注意的五个原则。虽然有时我们会违反其中的一两个原则，但是编写题项时我们仍然要认真对待。

1. 每一项应当表述且只表述一个意思。当一个题项表达多个意思时，受访者就会很困惑。他们会发现他们对这个题项中的每一不同观点的答案是完全不同的。试想如下一个题项：

我的指导老师充满活力，做事也井井有条。

这个题项问到关于老师的两个不同特征——有活力和做

事有条理。如果某人的老师只表现出其中的一个特点，那他/她该如何反馈情况呢？实际上，这些特点也许呈现互斥的关系。许多精力充沛、充满活力的人是相当懒散、毫无章法条理的。而做事严谨有条理的人往往不那么外向、充满朝气。一个受访者对某个只具备某一方面特长的老师进行评价时，往往踌躇不决，他/她也许会：(1)同意该项的表述，因为老师充分展现出了某一品质；(2)给出一个中等评价，对这两个特征的回答进行一下平均；(3)强烈反对该项表述，因为该项陈述的是一个老师同时具备这两项品质，但是该老师实际上只具备其中的一项而已。很有可能的是，受访者会在如何回答这个题项时产生分化，从而导致这个题项无效。对于每一个题项，请细致考虑它是否包含且只包含一个观点。两个观点的话，一定要分成两个题项来书写。

2. 同时采用正反表述。降低偏差的一个办法是将题项从正反两个不同角度进行编写。假设一个量表问人们对某事物的看法，那么有些题项就应当写成赞同的形式，而另外一些题项则应当写成不赞同的形式。举例来说，要设计一个量表来测评人们关于福利的态度，某些题项应当从赞成的角度进行书写（比如说，福利制度给那些穷人提供了有价值的服务），而另外一些题项则应该从不赞同的角度进行陈述（比如说，福利制度应当为我们许多的社会问题负责）。持赞成态度的人会同意头一个题项，而反对

后一个题项。持不赞同态度的人的回答则会与之正好相反。

通过变换问话的方向,因反馈倾向而导致的偏差能够达到最小化。比如默认(acquiescence)便是这样一种倾向——反馈者不论题项的内容如何都倾向于同意(或不同意)。表现出这种倾向的人会倾向于同时同意(或反对)上面两种表述,而不管其具体内容是什么。如果所有的题项都采用同一方向书写,那么有默认倾向受访者的量表分值将会很极端——要么非常高,要么非常低。他们的极端化评分将会扭曲均值的估计,以及利用量表分值做统计检验结果。如果题项利用正反措辞的数量相同的话,有默认倾向的人将倾向于得到中间值。它们的分值对均值的估计和统计检验结果的伤害将小得多。

在正反两种题项都存在的情形下,默认就会变得显而易见。我们能够分别计算出同一方向的题项的得分。每一个反馈者将会有正面表述的题项的得分,以及反面表述的题项的得分。如果两个相反措辞的题项的得分都很高,或很低,那么非常有可能是受访者带有默认倾向。然而此处需要提醒的是,默认倾向并不总是成为加总量表的一个问题(参见 Rorer, 1965; Spector, 1987)。

3. 避免使用俗语(colloquialisms)、惯用语(expressions)和行话(jargon)。最好使用简明的普通英语(plain English)(或者这个语言平时使用的形式),避免使用将把量表限定

为某个特定人群的术语。除非量表有特殊用途，否则量表的用语最好尽可能保持通俗易懂。甚至某些众所周知的惯用表达也会限制量表的使用对象和使用时段。比如，一个美式英语惯用语也许在其他的英语地区——比如英国或澳大利亚——很难理解。如果将量表译成另一种语言，问题将会变得更糟糕。

我们也应当考虑词汇会随着时间改变其意思和内涵。惯用语的意思也许会特别受到时间的限制。此处以设计一个关于堕胎选择的量表为例来进行一下考量。比如“我支持生命权立场”(pro-life position)这一题项，会被大多数人理解为反对人工流产(anti-abortion)。也许10年或20年后一种新的生命权运动跟堕胎没有任何关系，而只有极少数人会将生命权与反对堕胎联系起来。除非量表特别关注生命权运动这一用语本身的观点，否则最好采用一个更广为接受的用语。比如可以这么说:“我认为堕胎应该是非法的。”

4. 将受访者的阅读能力考虑在内。与前一原则相关联的是，受访者应当能够阅读并理解题项。得确保量表的可读性和词汇适合受访者。一个为大学生设计的量表，也许不适合用于调查高中生，仅仅因为词汇难度太高。另外也考虑题项的复杂性。受过高等教育的人群也许对题项中的复杂、抽象概念感觉很舒服，但那些概念对于那些受教育程度较低的人群来说也许就理解困难了。不幸的是，

多数受访者并不会就此提出抱怨。相反，他们会尽其所能做完量表，自然当他们碰到不理解题项时就会产生误差和偏差。要记得，如果量表语言越简单越基础，那么能够提供优良数据的适用人群就会越广泛。

对于那些没有阅读能力的人，我们可以口头将量表读给他们。但是，这样的做法必须十分谨慎。需要单独准备一份口语版本的量表。不能假定口语版本和书面版本具有同样的心理测量学上的特征。至少，应当找一个采用口语版本量表的样本进行题项分析。

5. 避免使用否定词来否定某一题项中的用词。通常我们会通过加一否定词，如“不”或“非”，来表述一个用词的反面意思。“我满意我的工作”这一肯定性的题项，可以通过加上“不”变成否定性的题项：

我不满意我的工作。

否定词的麻烦在于它们太容易被受访者忽略。在处理存在这类题项的量表时，我注意到许多人似乎都会误读否定性的题项。换句话说，他们常将否定性题项当成肯定性陈述来回答。

否定词被忽略会完全改变题项的意思，从而导致回答朝着完全错误的方向发展。当然，这种误差的存在也正是需要采用多题项的理由。单个错误只能轻微影响到多题

项的总分。但是，这些误差的确减小了量表的信度。

编写题项时不使用否定词，通常是很容易的。比如，上面的例子可以换个词来说：

我讨厌我的工作。

当人们读到这个题项时，他们就不太可能会误解它的意思了。

第4节 | 填表指南

量表编写的最后一件事可能就是填表指南了。指南可以实现两个主要目的。首先,可以给受访者如何使用量表提供指导。这对于许多的受访群体来说并无太大必要,比如大学生,他们已经习惯了填写这类量表。但对于那些对评分加总量表不熟悉的人来说,是很有必要的,因为他们并不清楚应该如何处理这个量表。表4.2就是工作控制点量表(WLCS)填写指南的一个范例。

表4.2 工作控制点量表(WLCS)填表指南

下面的问题是关于人们对工作和职业的看法和信念。这些问题提到的工作指一般性的工作,而不是您目前正在从事或曾经从事的工作。这些问题问的是您的个人观点,因此答案没有对错之分。不管您如何回答每个问题,我们能确保许多人也会如您一样回答这个问题。

请您对于每个问题标出同意或不同意。您可通过对最能代表您观点的选项画圈来表示。如果你非常不同意,则圈1;如果你一般程度地不同意,则圈2;如果有点不同意,则圈3。相反,如果你非常同意,则圈6;如果你一般程度地同意,则圈5;如果你有点同意,则圈4。

再次提醒您:您回答与工作相关的问题时,注意这里的工作指一般性的工作,而不是某个特定的工作。

第二类指南针对某些特别的概念。关于有些判断性工作,有时需要给受访者一些指导。例如,指南可以告知他们题项指向谁或指称某事。当某类量表测量民众对政治家的态度时,指南就应当标明哪类政治家在考虑范围内。

指南也可以给受访者提供一个一般的参考框架。在这种情形下,可以描述在程度或比例上明显较高或较低的人或物。比如,在一个工作自主性(job-autonomy)的量表中,可以提出一个大学教授的工作自主性非常高,而一个工厂工人的工作自主性程度非常低。可以具体阐述教授的工作比较自由,允许他们几乎完全能自己做主。相反,工厂工人则被描述为他们工作的各方面都被牢牢限制。这两个极端的例子给所有的受访者一个较为宽泛的参考基准,在此基础上他们对自己的工作作出判断。当然,个人有时也会从他们自身的特殊参考框架来解释这些描述。填写指南旨在尽量降低这种特殊参考框架,从而降低误差。

答案选项也可以界定得更细一些。例如,频次量表中的频率从“极少”到“非常经常”,这种情形下可以注明:每天一次的频率可以认为是“非常经常”,而每月一次可以认为是“极少”。

第 5 节 | 设计工作控制点量表(WLCS)

工作控制点这一概念是指人们对工作控制的看法。设计一个量表,让受访者对题项表示同意与否,似乎是最为合适的。我们选择了一个 6 点量表,其中同意方面提供 3 个选择,不同意方面也提供 3 个选择。表 1.1 正是这种量表,其中包括了答案选项,以及最终的 16 个题项。

起初的题项有 49 个:21 个关于内控型方向,28 个关于外控型方向。最终题项缩减为 16 个,该过程将在下一章讨论。

填写指南(见表 4.2)则指出题项中所指称的工作是一般性的,而不指具体某种或某个工作。这点非常重要,因为工作控制点被概念化为受访者的一个特征,而不是他/她的具体工作。虽然我们不能保证受访者回答问题时一定会使用这个参考框架,但是我们可将"一般性的工作"这一词重点标出,并提示两遍。并且,许多题项的用语都是指向一般性的工作。但是,有些题项也许会被理解为指具体的某项工作。

在使用量表时也需要用到指南。对于参加量表初测的大学生样本来说，这或许并无必要，但是当量表要适用于更为广泛的工作人群时，指南就应当加进去。毫无疑问，当请求某些受访者将来完成某个量表时，指南就非常必要了。

表1.1中一半的题项是关于外控型方向的，另一半的题项是关于内控型方向的。要计算整个量表的分值，其中半数的题项值必须反向转化。转换之后的得分为全部16个题项的总分。由于6个选项的赋值为1到6，总分也就从16(1×16)到96(6×16)。这里与罗特(Rotter，1966)的一般控制点量表一致，高分表示外控型，而低分表示内控型。

第5章

开展题项分析

量表构建的下一个步骤便要求收集数据，以便进行题项分析，其目的是得到一个试验性的量表版本——一个可用于效度验证的量表。如果前述步骤得以认真开展，并且还算幸运的话，这一步就只需要一次便可完成。否则的话，那些初始的量表题项备选库不能合成一个内在一致的量表，那就需要重复前面的步骤了。这也许包括重新界定概念特征，或者增写一些量表题项。

为进行这一步，我们必须找一个受访者样本来填写量表。如果样本的受访者尽可能代表最终的量表适用总体，那当然非常好。但这总是不可能的，许多的量表测试初期总是用在大学生身上，因为他们比较容易找得到。在这样的情形下，将量表应用到没有接受高等教育的人群时就必须特别小心。很有可能新的样本群体反馈不同，只是因为量表的阅读难度太高了。

题项分析需要一个大概包括 100 至 200 个受访者的样本。WLCS 的初始样本是 149 个。这种规模的样本通常在

大学中比较容易获得。在其他的环境中，或要求某些特殊的总体时，得到这么一个数量的受访者样本或许是不容易的。量表设计的后续阶段需要多得多的受访者。

数据分析所涉及的统计的复杂程度不会超过计算相关系数的复杂程度，但是如果人工计算的话非常耗时。SPSS-X这一统计软件含有一个题项分析的固定程序，它能进行本章讨论的题项分析。在大型计算机和微机上我们都能很容易找到这个程序。本章后面会有些关于程序应用的指导。

第1节 | 题项分析

题项分析的目的是找出那些能够形成一个内在一致性量表的题项，剔除那些不一致的题项。内在一致性是题项的一个可测特征，意味着所有的题项都测量同一个概念。它反映各题项相互间的相关程度。无相关性则意味着那些题项并不代表一个共同的潜在概念。整个题项间的内在一致性表明它们存在相同的方差（common variance)，或者说它们是同一个潜在概念的指标。该概念的内在本质当然都是有待质证的。

题项分析为每一单个的题项与其他题项的相关程度提供信息。这可以通过每个题项计算得出的余项系数(item-remainder coefficient)来反映。这个统计量也被称为部分—整体或者题项—总体相关系数(part-whole or item-whole coefficient)。余项系数是每个题项与剩余题项的总和之间的关系。对于一个有10个题项的量表来说，题项1的余项系数是通过计算题项1的分值与题项2至10的总得分之间的相关系数来测量。题项2的余项系数则是通过

计算题项2的得分以及题项1、3至10的总得分的相关关系来测量。这样的题项分析要对所有的10个题项进行一遍计算。

如果所有的题项并非按照同一方向进行措辞,也就是说有些题项是正面措辞,而有些题项是负面措辞,那么负面措辞的题项的得分必须进行反向计算。否则对正面措辞的题项的回答将会抵消对负面措辞的题项的回答,然后绝大多数的受访者都只有一个中等分值。对每一个题项来说,高分值应当代表某个概念的高水平度呈现,而一项低分值应当代表某个概念的低水平度呈现。因此在某概念上高水平度的受访者将会表现出同意正面措辞的题项,不同意负面措辞的题项。为了准确地对他们进行评分,同意的程度应当与反对的程度均等。对于一个六点量表来说,对正面措辞的题项的同意6分应当等同于负面措辞的题项的不同意1分。举例来说,强烈同意下面的表述:

我喜爱牛奶。

大约等同于强烈反对另一个表述:

我讨厌牛奶。

在六点量表中,强烈同意前一个表述,会给受访者6分,强

烈不同意第二个表述也会给 6 分。（参见第 4 章关于题项反向计分的阐述。）

题项分析为每一个题项提供了一个余项系数。那些余项系数最大的题项将会得以保留。这里有几个办法来决定哪些题项需要保留。如果决定该量表得有 m 个题项，那么余项系数最大的 m 个题项可以被选。或者采用另一种办法，设定一个余项系数标准（比如 0.40），然后保留所有大于该系数标准的题项。这两种方法可以同时采用，即保留最多 m 个题项，并且它们都符合设定的某个最低系数值。

在题项的数量与余项系数的大小之间也存在某种平衡。题项越多，余项系数越低依然能够获得一个好的、内在一致的量表。内在一致性还包括另一个统计量——系数 α（coefficient alpha）。

系数 α（Cronbach, 1951）是量表内在一致性的测量值。它是题项数量与它们之间的相关程度的一个直接函数。题项数量的增加，或者题项间相关程度的提升，都可提高系数 α。假如题项数量足够多的话，即便它们之间的相关程度很低，这些题项也可获得一个相对高的系数 α。

探讨其中的原因，我们还得回到经典检验理论。如果假定所有的题项都指向某个单一的潜在概念，题项之间的相关系数则代表误差的倒数。换句话说，每个题项中与其他题项无关的部分，可假定是由误差造成的。如果误差相

对较小，那题项之间的相关就高。将误差均化为接近0，并不需要太多的量表题项数。另一方面，如果相关很小，那误差就很大。这种大的误差也会被平均掉，但这需要数量较多的量表题项。

不过，请别忘了，经典检验理论所提供的几种假设在现实中也许有效，也许无效。低相关度的题项也许可以形成一个内在一致的量表，但这并不必然保证这些题项就能反映出一个单一的潜在概念。如果我们利用两个量表来测量相关但不同的概念，它们所有题项的合并也会得到内在一致性，可它们反映的其实是两个不同的概念。由题项分析所得出的统计量是选择题项的好指南，但是题项的内容必须经过仔细审查，以便得出有关这些题项测量内容的结论。

系数 α 反映的是内在一致性的信度，但它并不必然反映随时间变化的信度(reliability over time)。测量某些概念(比如情绪)的量表也许具有内在一致性，但是随着时间的变化它的信度会降低，因为概念本身会随着时间的变化而发生变化。

系数 α 的作用看上去像相关系数一样，但是系数 α 并不是个相关系数。它通常是从0到1.0之间的正值，数值越大表明内在一致性程度越高。农纳利(Nunnally, 1978)提供了一个广受认可的经验法则，即一个量表要证明其具有内在一致性，α 值应至少不小于0.70。许多量表的 α 值

不能达到这个水准，导致其应用性受到质疑。如果题项之间呈负相关，α 值也可能为负数。如果题项的分值没有得到正确的反向转化的话，就会发生这样的情况。假定所有的题项分值的方向都正确的话，那系数 α 应该就是正值。

系数 α 包括将量表总得分的方差（所有题项分值的加总）与单个题项得分的方差的比较。从数学上来说，如果题项之间不相关，量表总方差将会等于每一题项方差的总和。

当题项之间的相关度越来越高，量表总方差将会上升。例如，假设某个量表有三个相互独立的题项，其总方差就是三个题项的方差总和。如果每个题项的方差值为1.0，那量表的总方差值就是3.0。如果题项之间相关，那量表总方差就会大于3.0，即使单个题项的方差仍为1.0。

系数 α 的计算公式为：

$$\alpha = \frac{k}{k-1} \times \frac{s_T^2 - \sum s_I^2}{s_T^2}$$

其中，s_T^2 是题项总和的总方差，s_I^2 是单个题项的方差，k 是题项的数量。可以看出，该方程的分子包含着量表总方差和各题项方差之和的差。这两种方差之差除以量表总方差，得出一个比值系数。最后，该比值系数乘以一个题项数量的函数就得到系数 α。

在选择量表题项时，我们同时使用余项系数和系数 α。

选择过程包含一系列步骤，比如删除一些题项、检验 α 值，再剔除更多题项，再进一步验证 α 值，直到最后选出一组题项。删除“坏”题项会提高 α 值，但同时减少题项数又会降低 α 值。删除许多弱题项会不会提高 α 值，取决于有多少题项留下来，以及被删除的题项有多弱。

表 5.1 使用题项分析来选择题项的例证

步骤	题项	余项系数	去掉该题项的 α 值
1	1	0.53	0.68
	2	0.42	0.70
	3	0.36	0.71
	4	0.10	0.74
	5	0.07	0.75
	6	−0.41	0.80
	7	0.37	0.71
	8	0.11	0.79
	9	0.55	0.68
	10	0.42	0.70
		系数 $\alpha = 0.72$	
2	1	0.56	0.79
	2	0.43	0.81
	3	0.31	0.84
	7	0.39	0.82
	9	0.58	0.78
	10	0.44	0.81
		系数 $\alpha = 0.83$	
3	1	0.57	0.79
	2	0.44	0.80
	7	0.40	0.80
	9	0.59	0.79
	10	0.45	0.80
		系数 $\alpha = 0.84$	

表 5.1 是个虚拟的筛选题项的过程，其中有 10 个题项已应用到一个样本中。步骤 1 标明了每个题项的余项系数，从 −0.41（第 6 项）到 0.55（第 9 项）。在这些相关系数下方是量表的总体系数 α 值（此处为 0.72）。另外，最后一栏是指该项被剔除时所获得的 α 值。这一栏很有帮助，因为它可以显示对于每个题项，删除它对整个量表内在一致性有多大影响。

该表中有 6 项的余项系数大于 0.35，并且任一项如果被移除的话都会降低系数 α。剩余的 4 项，如果被移除，则 α 值会提高。有意思的是，第 6 个题项的系数 α 很高，但它是负数。也许是计分错误造成的，本该对此题项做反向计算。如果原因是这样的话，这个错误就可以得到修正并且重新对此做一遍分析。但是，也常有另外的情形发生：某题项看起来已经做过反向计算，却依然得到一个大的负系数。这种情形下，该题项就没有发挥本该发挥的作用。这里明显出错了，需要我们仔细检查这个题项。

无论何时出现大的负余项系数，我们都有必要仔细检查所有前面的步骤，确保前面的步骤没有出错，及整个量表设计的过程到此为止并没有出错。

余项系数出现负数时，首先应该考虑的事情是题项本身可能没有写好。一个题项常常开始显得很合理，但实际上很含混。第二个可能的问题是题项不适合当下的受访者。他们也许不能理解题项，也许没有足够的信息对题项

做出正确反应。如果题项本身和受访者都不是问题，那有可能是概念化过程存在不足。这时需要返回到概念化步骤，看看概念的界定是否正确。或许是概念本身没有效度，或许是概念的可操作化不正确。举例来说，假定某种个性特征会在几种行为中反映出来。但是，所获数据却显示，这些行为并不同时发生（至少通过受访者的报告显示如此）。这些数据就对正在考虑的概念的可行性提出质疑。

一些量表设计者依赖经验来决定题项的措辞方向。这是很有风险的，可能导致量表效度受到质疑。题项分析不应当用于决定题项计分的方向。存在一种例外就是当题项分析发现得分方向上的错误，就像前面讨论的那样。当得分方向转换时，题项会出现异常。当另一个题项的符号改变时，正的余项系数会变成负的。在所有的余项系数都变成正值之前，需要重复多次符号的反向转换。即便我们可以得到一个可接受的 α 值，量表的效度依然可能存疑，因为概念基础可能很弱，题项本身质量也很低，同时还存在题项指向多个概念的可能性。这个世界上并不存在设计好量表的捷径。

现在假定表5.1的题项6是个写得很差的题项，与题项4、5、8一起被剔除出去了。第二步显示的就是一个有所改善的量表。系数 α 提升至0.83，并且量表也更有效率，因为它将题项从10个精简为6个。这里请注意，其中5个题项的余项系数稍稍提高了。但题项3的余项系数从第一

步的 0.36 下降到第二步的 0.31，它的移除会将量表的系数 α 稍稍提至 0.84。在第三步中，题项 3 被剔除了，整个量表的题项减少至 5 项，α 值为 0.84。注意，剩下 5 个题项中任意一个题项的移除，都会降低量表的 α 值。

这里存在一个令人费解的结果，与题项 3 有关。看上去令人奇怪的是，当第二步中那些“坏题项”被移除时，题项 3 的余项系数下降了。由于题项之间相关关系的复杂性，这种情形是可能产生的。虽然该题项与最终剩下的 5 个题项都存在相关，但是它也与一些被删除的题项相关。甚至，它与被删除题项的相关要高于它与保留题项的相关。因此，当这些题项删除时，题项 3 对量表整体的贡献也就降低了。最终，删除题项 3 提升了整个量表的质量。

这里关于系数 α 的讨论，就简要结束了。如欲继续学习，请有兴趣的读者阅读另外的心理测量学文献（例如，Allen and Yen，1979；Nunnally，1978）。

第2节 | 题项选择的外在标准

内在一致性是选择题项最常用的标准。一个替代性的方法是根据题项与某些外在标准的相关关系进行选择(或剔除),它有时也与内在一致性标准配合使用。换句话说,当题项与某个感兴趣的变量相关时,它们会被保留,或者被剔除。

当有关偏差的问题出现时,一个办法就是独立测量偏差,然后剔除与此偏差有关的题项。在实际操作上,这可能包括将量表进行一次样本测试,测量同一样本人群中的偏差变量。每一个题项都和有偏差的变量存在相关。假定各题项与偏差变量的相关程度都有所不同,就只有那些相关性小或者无关的题项可被保留。

题项选择的一个普遍外在原则是社会期待(social desirability,简称 SD)(Crowne and Marlowe, 1964)。从量表设计的角度来看,SD 反映的是个人从一个社会期待的角度回答量表的倾向。表现出高 SD 的人倾向于同意那些对他们自身有利的表述(比如,“在别人有困难时,我毫不犹豫

会伸出援助之手”)，倾向于反对那些不利于自身的表述(比如，“我能记得通过‘装病’来推脱某些事情”)。这两个题项的例子都来自克罗恩—马洛社会期望量表(Crowne-Marlowe SD scale)(Crowne and Marlowe，1964)。

设计中的量表的每一题项都能与SD值存在相关。与SD值存在显著相关的题项应当从最后量表中被删除。通过这种方法，我们可以得到一个无SD偏差的量表。即，对量表的回答将不受受访者SD倾向的影响。当然，对于某些概念，不太可能写出独立于SD的题项。这种情况可能发生是因为概念本身就与SD的潜在概念存在相关。这时，量表的效度就有待质疑了。该量表到底是测量某个要测的概念还是SD?

如果事物进展顺利，一旦题项选择工作完成，我们就获得了一个可接受的、内在一致的量表。该量表的试验版本现在就准备进入下一步设计程序了。我们还需要做些额外的工作，对把量表应用到第二个样本得到的结果重新进行题项分析，进一步建立信度和效度。这些工作可以同步完成。

第3节 | 量表的进一步完善

每个量表的初次设计并不能保证其能获得充分的内在一致性。也许有几个题项符合保留标准，但是系数 α 可能太小。如果发生这种情况，量表需要增写一些题项，需要收集更多的数据，题项分析也需要重做。当初始题项数量太少，或太多题项质量不高时，上述情形就会发生。如果概念构建工作较差时，这样的情形也会发生。

当未达到可接受的内在一致性水准时，首先的问题是确认原因是否与概念定义有关还是与题项有关。这个问题很难确定回答。我们通过重新检视那些保留题项和删除题项，可以得到一些线索。当待测的概念定义太过宽泛或模糊时，这可能非常明显。例如，假定某一单维度概念含有四个相关要素，待写题项来反映其中的每一个要素。如果反映该概念每一要素的题项之间不存在相关，那么题项分析只能留下的题项太少可能无法得到一个内在一致性的量表。也许概念要比预想的更狭隘，那么应该只集中反映四个要素中的一个要素。相反，如果概念是多维度

的，其中的每一个要素都代表一个独立的方面，那么这种情况则需要设计多个子量表。

仔细检查被删除的题项，也许会发现它们仅仅是因为质量不高而已。它们也许含有第 4 章讨论的一些问题。或者，它们不适合于受访群体。探究这些可能性的一个方法是，找几个受访者，问问他们对于每个题项的反应和解释。

当题项分析导致题项太少时，估计一下得增加多少题项，以获取一个可接受的内在一致性标准将很有帮助。斯皮尔曼—布朗预测公式（Spearman-Brown prophesy formula）可用于计算需要增加的题项数量，以获得给定的内在一致性水平。给定某一特定数量的题项的系数 α，预测公式可以表明题项增加或减少对系数 α 的影响。这个公式的计算是基于这么一个假设，即增加或减少的题项与初始的或保留的题项的质量水准是相当的。否则，该公式将会高估或低估所需题项的数量。

如果题项的数量是基于某一特定因子来增加或减少的，那么预测公式可以估计量表可得的系数 α。比如，如果题项数量翻倍或减半，那么系数 α 会如何变化呢？该公式可以回过头来去说明为获得某一给定系数 α 值，需要增加或减少多少题项，即量表该有多长。

该预测公式为：

$$r_N = \frac{k \times r_O}{1 + (k-1) \times r_O}$$

其中，r_N 是新量表的信度（系数 α），r_O 是原量表的信度，k 是题项增减的因子（比如，2 表示长度翻倍，0.5 表示减半）。

表 5.2 展示了题项增减 1/2、1/3、1/4、2、3 或 4 倍时的效果。其中几个代表性的 α 值（从 0.50 到 0.90）也列出来了。从这个表格可以看出，一个系数 α 值为 0.70 的量表，如果题项数量翻倍，系数 α 将增至 0.82；如果题项数量减半，系数 α 将降至 0.54。

表 5.2 基于斯皮尔曼—布朗预测公式（the Spearman-Brown Prophesy Formula）进行增减题项的函数所得的信度

初始 α	变化的尺度（因子数）					
	0.25	0.33	0.50	2	3	4
0.50	0.20	0.25	0.33	0.67	0.75	0.80
0.60	0.25	0.33	0.43	0.75	0.82	0.86
0.70	0.37	0.44	0.54	0.82	0.88	0.90
0.80	0.50	0.57	0.67	0.89	0.92	0.94
0.90	0.69	0.75	0.82	0.95	0.96	0.97

预测公式能够为最终的量表提供一个题项的目标数目。比如说，一个量表初始的题项数目是 15 个，经过题项分析减少到 5 个，此时的系数 α 为 0.60。通过预测公式（参见表 5.2）可知，要达到系数 α 为 0.75 的话，题项数目必须翻番；要达到系数 α 为 0.82 的话，题项数目必须是原来的三倍。现在假定新的题项在质量上与原题项一样，并且只有三分之一的题项被保留了，还需要增加 15 至 30 个题项。当然，对保留题项和剔除题项的仔细检查和分析，有助于

提高第二次写题项的质量。大部分第二次写的题项应该被保留。

预测公式也能被用于减少题项。假设现在有一个量表,有 30 个题项,系数 α 为 0.90。该量表或许太长,考虑到其具有很高的内在一致性,它应该可以被缩减得更短一点。表 5.2 显示,如果将题项减少一半(至 15 个),内在一致性系数 α 可以控制在 0.82。我们再次假设保留题项和删除题项具有均等的信度。如果题项分析数据可得的话,那我们将会保留 15 个最好的题项,这样 0.82 这一系数将会是我们可以得到的系数 α 的保险底数。

第4节 | 多维度量表

到目前为止，我们所有的讨论都集中于单维度量表的设计。这种量表构建于具有相对同质性、单维度概念的基础之上。然而，许多概念其实是相当宽泛的，包含多个方面或多个维度。对于复杂事物（比如对于政府、工作、家庭和生活等）的态度，就包含多个方面。为了充分表达个人的态度和意见，人们需要多个子量表来反映多个维度。

多维量表的设计与单维量表的设计也没有多大不同。在概念化的最后，需要将各个不同部分具体化。这些组成部分不一定非得不相关，并且多维量表的子量表之间常常是相关的。但是，从概念上说，它们应是相互独立区分的。

多维量表的设计步骤也可遵循前述步骤，每一子量表与其他子量表可并行设计。针对不同的预测目标编写题项。所有的题项通常都混合在一个初始题项备选库（initial item pool）中。它们会被试用于同一个受访群体。每一子量表都单独进行题项分析，题项分析只针对某一子量表的题项。通常所有的题项总合起来作分析，得出一个总分，

但只有当概念上有意义时才会这么做。

对于多维量表来说，如果每一个题项是且仅是某一子量表的组成部分，那么这将是最好的情形。这么说，有如下几个理由。首先，在设计子量表时，我们就假定某一概念非常复杂，必须要被几个概念上独立的部分来代表。如果各个组成部分的内容相互重叠，以至那些子量表必须共享题项，那么很可能它们就不是独立的，不足以保证区分不同的部分。

从测量的角度来说，共享的题项会导致一些问题。其中最为重要的可能是，反映在子量表中各构成要素之间的比较会变得模糊。对于概念之间潜在关系的推断也将变得不可能。换句话说，子量表之间的相关是因为它们的潜在概念相关，还是因为它们共享某些题项，无法判断。同样，当不同的子量表与另一标准变量(criterion variable)相关时，将不可能把这种类似的关系归因于其潜在概念或题项重合。

当子量表的题项含有虽然不同但相似的内容时，也会产生问题。由于两个概念分享共同的要素，题项内容就会产生重合。以愤怒和焦虑这两种情绪状态为例。人们能够区分这两种情绪，但是它们也有一些共同的特征。它们都含有一种令人厌恶的反应。比如“我感觉紧张”“我感觉很差”“我感觉不舒服”等这些题项都可能用于反映这两种情绪。

如果重合题项被剔除，那么量表的内容效度也将会打折扣，因为它将可能不反映所测概念的全部范围。如果重合题项被保留，那么对于量表间关系的解释将会变得很危险。这是因为每一个量表都含有可以反映另一个概念的题项。如果一个受访者处于高度焦虑状态，但不是愤怒，那么他/她将会在测量愤怒的量表中的重合题项上获得高分。与愤怒相关的分值将会因为焦虑而变高。接着两个量表之间的相关将在一定程度上增加，就是因为受访者仅仅在一个概念上得分很高。如果焦虑和愤怒之间的相关性存在的话，那么我们必须小心不要误读了其中的原因。是测量对象因为经历了某一种情绪，就会倾向于经历另一种情绪，还是仅仅因为某种程度上每张量表都反映另一种情绪？对于这种问题，可不好寻找答案。

在上述例子中，我们也许可以期待焦虑和愤怒能够区分得足够清楚，题项重合的情形可以避免。也许我们可以请求受访者对他们焦虑和愤怒的程度进行打分，假定愤怒而非焦虑(nonanxious angry)的人不会报告对于焦虑的经历，焦虑而非愤怒(nonangry anxious)的人不会报告对于愤怒的经历。其他的概念也许并不能如此容易区分。尼科尔斯等人(Nicholls et al., 1982)曾讨论在概念层面上的重合对数据解释的影响是如何不可避免的，是如何成为问题的。

在量表的概念构建过程中，我们应当仔细考量概念的

重合之处在哪里，区分之处又在哪里。构建一个新的概念时，其中的部分工作就是要搞清楚它与之前存在的概念有什么差别。在量表构建时，应当努力避免新量表与其他量表内容上的重合，除非因为要测量的概念本身有重合，而导致内容上的重合。当量表共享题项内容时，必须谨慎处理对它们相互关系的解释。

第5节 | 利用 SPSS-X 执行题项分析

利用计算机来进行题项分析,虽然不是绝对必要,但我们还是强烈推荐。即便题项很少,但如果采用手工计算,题项分析还是要花费许多时间。在计算机和合适的软件普遍可获得的情况下,如果我们还放弃使用计算机的话,那真是有点犯傻。

目前最为流行的能用于题项分析的统计软件包之一是 SPSS-X(SPSS Inc., 1988)。该软件包的大型主机版本和微机版本,至少其中一种可以在几乎所有的美国大学里找到。"信度"程序能用于执行题项分析。该程序可以计算题项统计量和前面讨论的系数 α,也能够进行更为复杂的运算,本书就不作深入讨论了。

表 5.3 就是一些用于 WLCS 的题项分析的程序。其中,在大型主机上运行这个分析的系统命令被忽略了,因为这些命令因不同的计算机会有不同。这里也需要提醒的是,计算机软件会随时更新。也许这里列出的程序在将来的计算机版本上就不能运行了。

表 5.3 执行 SPSS-X 题项分析的程序

```
DATA LIST FILE'A: WLCS.DAT' FIXED
    /ID 1-3 S1 TO S49 4-52.
  COMPUTE S1 = 7- S1.
  COMPUTE S2 = 7- S2.
  COMPUTE S3 = 7- S3.
  COMPUTE S4 = 7- S4.
  COMPUTE S7 = 7- S7.
  COMPUTE S12 = 7- S12.
  COMPUTE S15 = 7- S15.
  COMPUTE S16 = 7- S16.
             ⋮
RELIABILITY
    /VARIABLES S1 TO S49
    /SCALE(WLCS) S1 TO S49
    /MODEL ALPHA
    /SUMMARY TOTAL.
```

这个例子假定,对于每个受访者,WLCS 最初的 49 个题项都被输入到一个单一的数据行(data line)上。每一行数据,代表一个单一的受访者,有一个受访者的 ID(identification number),分布在第 1 列至第 3 列。这 49 个题项根据顺序列在第 4 列至第 52 列,命名为 X1 至 X49。

程序的第一行表示数据所在的位置。这个例子采用的是与 IBM 兼容的微机版本程序。数据存在 A 盘,文件名为"A: WLCS.DAT"。程序的第二行显示变量名和它们所在每条数据的位置。接下来的八行是对适当的题项进行反向编码的计算语句。这里只显示了头八行。下一行的语句提出信度程序。在这个语句下面列出了所有题项的变量语句,不管是否所有的题项都用于题项分析。这里列

出了 X1 到 X49。下一行确定了分析类型，命名为 scale (WLCS)，并列出选择的题项。这个例子选择了所有题项，虽然也可选择更少一点的题项。对于多维度量表来说，这条命令指定某个特定子量表的题项。下一行是模型，在这个例子中是系数 α。这告诉程序去执行这里讨论的题项分析。接下来的两行是对子量表进行题项分析的语句，将重复列出。每两行列出对应于一个子量表的不同的题项子集。最后一行列出了要计算的统计量。这有许多可能的选择，这在 SPSS-X 手册中都有详细介绍(SPSS Inc.，1988)。

从这里能看出，利用计算机来做分析本身是很容易的。当题项逐步被剔除，程序中的变量和量表命令也随之改变，仅列出了保留的那些题项。

表 5.4 是利用微型计算机版本运行信度程序得到的输出结果的一个样本。为了节省篇幅，这里只列出了头 6 个题项和最后 6 个题项。输出结果包括前面讨论的特征。这个程序能生成许多附加的东西。这个程序首先列出了分析中包括的题项。这个表的下面是一个题项的总和统计量表。这个表包括对于每个题项，删除该项得到的总量表的均值和方差、余项[或题项—总和(item-total)]系数、系数 α。在这个表的底部是系数 α。(注意：这是微型电脑版本。大型主机版本会给出更多附加信息，这里都没有讨论。)

表 5.4 工作控制点量表(WLCS)的 SPSS-X 项分析输出结果

Reliability Analysis—Scale(WLCS49)

1.	S1
2.	S2
3.	S3
4.	S4
5.	S5
6.	S6
⋮	⋮
44.	S44
45.	S45
46.	S46
47.	S47
48.	S48
49.	S49

RELIABILITY ANALYSIS—SCALE(WLCS49)

Item-Total Statistics

	Scale Mean If Item Deleted	Scale Variance If Item Deleted	Corrected Item—Total Correlation	Alpha If Item Deleted
S1	137.463 1	418.507 1	0.297 5	0.875 3
S2	138.691 3	418.795 9	0.270 3	0.875 8
S3	138.892 6	419.650 6	0.301 3	0.875 2
S4	138.926 2	415.068 8	0.383 5	0.873 9
S5	138.986 6	419.297 1	0.249 3	0.876 2
S6	139.543 6	415.317 3	0.370 1	0.874 1
⋮	⋮	⋮	⋮	⋮
S44	138.422 8	420.840 3	0.275 6	0.875 5
S45	138.127 5	412.855 3	0.452 4	0.872 9
S46	139.302 0	417.104 1	0.514 3	0.873 0
S47	139.973 2	425.715 5	0.235 7	0.876 0
S48	138.335 6	420.062 3	0.212 4	0.877 1
S49	139.067 1	417.833 3	0.439 0	0.873 6

Reliability Coefficients

N of Cases = 149.0 *N* of Items = 49

Alpha=0.876 9

第6节 | WLCS 的题项分析

WLCS 的最初的题项分析包括 49 个题项。这个量表被应用到 149 个本科生中，这些学生构成了初始量表设计样本。因为控制点是一个个人特征变量，所以有人担心社会期望可能引起的偏差效应。这个问卷包括了克罗恩—马洛社会期望量表（Crowne-Marlowe SD scale）（Crowne and Marlowe，1964），因此这个量表能作为题项选择的一个外在标准。而且，问卷包括几个量表和能用于效度验证的题项，并假定这些题项构成了一个内在一致性的量表。这些将在第 6 章讨论。

题项被分成内控和外控型两个方向。最初的计划是让每种题项类型在量表中的数量相当。采用这种方法的理由，是尽可能地抵销由措辞方向引起的偏差。

所有这 49 个题项的系数 α 是 0.86，这个值是在可接受的范围内。然而，一个包括 49 个题项的量表比理想的量表要长得多。考虑到其中 28 个题项的余项系数大于 0.295，量表的长度能大大地减少。最终的量表选择了最合适的 8

项外控型题项和8项内控型题项。所有外控型题项的余项系数都比内控型题项的大。这也许表明我们能删除内控型题项。这些内控型题项没有被删除是因为初始的计划是要均衡两种题项类型的。因此,如果计算得到的系数 α 是可接受的,那么就把这两类题项都包含在量表中。

将这49个题项的每一项和社会期望量表之间的相关都计算之后,我们决定删除任何与SD在统计上存在显著相关的题项。但是,量表中没有一个题项基于这个原因必须删除。

表5.5包括最终量表的16个题项的余项系数。表里

表5.5 工作控制点量表(WLCS)的题项分析

题项数	措辞方向	初始余项系数	最后余项系数
1	内控型	0.35	0.29
2	内控型	0.34	0.30
3	内控型	0.41	0.35
4	内控型	0.30	0.24
5	外控型	0.52	0.52
6	外控型	0.56	0.59
7	内控型	0.31	0.32
8	外控型	0.55	0.50
9	外控型	0.62	0.64
10	外控型	0.62	0.68
11	外控型	0.52	0.54
12	内控型	0.34	0.34
13	外控型	0.59	0.64
14	外控型	0.50	0.56
15	内控型	0.48	0.46
16	内控型	0.44	0.43

列出了每个题项的措辞方向，对 49 个题项进行题项分析得出的余项系数和对最终 16 个题项进行题项分析的余项系数。在对所有 49 个题项的初始题项分析中，余项系数的取值范围从 0.30 到 0.62。当对最终的 16 个题项重新进行题项分析时，余项系数稍稍有些变化。对于内控型措辞的题项，其余项系数倾向于变小；对于外控型措辞的题项，其余项系数倾向于变大。最终的余项系数的取值范围从 0.24 到 0.68。

虽然去掉两个题项实际上能稍微改善量表的系数 α，但是最终的量表保留了这两个题项，这是为了均衡内控型和外控型题项。一个替代的策略应该是删除这两个题项（这两题项的措辞都是内控型的），同时也删除两个外控型题项。当这样做时，这个量表的系数 α 会变得更低。因此，保留这 16 个题项。整个量表的系数 α 是 0.845。这比包括 49 个题项的总量表的系数 α 要低一些，但仍在可接受的信度范围内。

这样，我们得到一个试验性的 WLCS 版本，其内在一致性在可接受性的范围内。社会期望，这个主要的潜在偏差来源看起来最小化了。这个量表看起来可以进入下一步的效度分析，但是内控型措辞题项的余项系数相对低仍然存在些麻烦。我们在第 6 章将继续讨论这个问题。

第6章

效　度

量表设计的最困难部分是效度——也就是说，解释量表分值代表什么。如果量表是内在一致的，那么这个量表肯定在测量某样东西，但是确定这样东西是什么是个很复杂的问题。这个困难部分源于效度仅仅存在于感兴趣的概念和其他概念之间假设关系的系统中。效度检验包括同时检验有关概念的假设和有关量表的假设。

如前面讨论所示，社会科学中的许多概念是在客观经验中不存在的理论抽象。那么，当一个概念自身不能直接被证实有效时，怎么才能有效地测量这个概念呢？全面回答这个问题将涉及哲学领域，超过了本书讨论的范围。

典型的量表效度验证策略包括在待测概念和其他概念的一系列假设关系的情形下，检验待测的量表。也就是说，建立有关这个概念的原因、效应和相关性的假设。量表用来检验这些假设。存在支持这些假设的经验意味着这个量表有效。

量表的效度验证就像理论的检验，因为其合适性不能

被证实。但我们可以收集经验证据来证实或反驳效度。当我们收集到足够多能证实效度的数据时，我们就可初步宣布构建的量表是有效的。量表的使用人员将采纳量表代表的东西的理论解释。当然，潜在的概念就是一个理论实体。概念本质的界定(the conception of the nature of the construct)及它与其他概念相关的原因都基于某个理论框架，这个理论框架将来可能被一个重新定义这个概念的新框架替代。

如同一个理论，一个概念能被初步采纳是因为它有用武之处。也就是说，一个有用的概念是它与其他概念之间的关系的理论体系的一部分，这个理论体系可以解释、预期和导致对我们感兴趣的现象的控制。经验的效度验证证据为从理论上预测待测概念和其他概念的关系提供了支持。这显示了概念的潜在用途。

直到题项分析已经执行且题项也选好后，效度验证的工作才能开始。当最初的题项备选库被应用到受访者时，我们就能收集有关效度验证的数据。假设早期的效度研究看起来效果不错，量表将可继续用于设计的研究中，以便(至少在某种程度上)提供效度数据。

第 1 节 | 研究效度的方法

这章将讨论三种不同的建立效度的方法。校标效度(Criterion-related validity)包括检验量表将如何与其他变量相关的假设。区别效度(Discriminant validity)和聚合效度(Convergent validity)常常放在一起来研究,包括比较研究变量之间的关系的强度和模式。这两种研究效度的方法包括研究待测量表和其他变量之间的假设关系。这里将包括用因子分析来探究量表维度。

校标效度。有好几类校标效度,所有的都包括把待测量表的分值和其他量表或标准(criteria)的分值进行比较。这样的比较包括研究待测量表的分值和其他变量的分值的相关性,也包括比较不同的可识别受访群体在待测量表上的差别。

校标效度研究从生成待测概念和其他概念之间关系的假设着手。通常情况下,量表是设计用来检验一个现有的成熟理论。在这种情况下,量表也能被用来证实理论生成的假设。在另外一些场合下,量表设计用来评价一个概

念，这个概念可能不是来自成熟的理论。在这种情况下，我们必须做理论工作，从而生成有关这个概念的假设。在任何一种情况下，校标效度必须基于假设。这里，最理想的情形是已有数据支持这些假设。当证实有关这些潜在概念的数据已存在时，有关效度的结论能更有把握地得出。

同时效度(Concurrent validity)能通过同时从一个受访样本回答待测量表和标准(这个标准假设与这个量表相关)收集的数据来检验。这里的同时指同时收集所有的数据。通常情况下，我们假设待测量表与一个或多个标准相关。量表和假设的变量之间存在统计上显著的相关关系将认为量表是有效的。

预测效度(Predictive validity)在很大程度上与同时效度一样，不过预测效度是在收集标准变量的数据之前收集待测量表数据。因此，这在检验一个量表能多准确地预测未来的变量。相对于同时效度，预测效度的优势在于，当量表可能在将来实际应用时，它更好地显示了一个量表能多好地预测未来的变量。例如，通过受访者的特征或态度来预测受访者是否退出某项活动(如辞职或退学)。

为了进行同时效度研究，常常把待测量表放在包括测量多个变量的问卷中。多重假设就通过这种策略来检验。当然，如果有多个假设，那很可能有一个或多个假设被拒绝。尤其当假设不是建立在坚实的理论基础上时，这种可能性将更大。在这种情况下，理论和量表效度验证的相互

依赖性表现得最明显。当至少某些假设——尤其是那些关键的假设——被接受时，量表设计人员常常宣称这个量表是有效的。他们断定（或许仅仅是希望）某些假设是错误的，而不是量表缺乏有效性。

这种情况需要一个正确判断，权衡每个假设对效度建立的重要程度。某些假设对于概念来说是如此重要，以至于如果这些假设被拒绝，那么量表或概念的效度将受损。当然，我们也无法排除经验检验自身存在瑕疵这种可能性。也有可能用来检验量表的数据无效，这些数据常常是根据其他量表生成。或许研究设计的偏差或混杂影响结果。也有可能因为样本量太小，检验缺乏足够的统计功效（statistical power）。

已知群体效度（Known-groups validity）基于某个受访群体的量表分值将高于其他受访群体的假设。这类效度和另两类效度的主要差别在于这类效度的标准变量是分类的，而不是连续的。待测量表的均值能在位于分类变量的每个分类中的受访者之间进行比较。虽然使用的统计量不是相关系数，但是均值差异仍然反映量表和标准分类变量之间的关系。

为了进行已知群体效度研究，假设必须指定某个群体的量表分值高于其他群体。例如，我们可以假设一个关于政治保守性的量表，共和党比民主党平均得分更高。同样地，我们也可以假设公司总裁比数据录入职员在测量工作

复杂程度的量表上得分更高。

下一步是识别受访群体，并让他们填写量表。然后从统计上比较这些群体的均值是否像假设的那样存在差异。T 检验或方差分析（这取决于比较群体的数量）将用来确定这些差异是否统计上显著。

受访群体的可获得性决定了能做的已知群体比较的本质。通常情况下，本该得分高或低的群体可能不容易获得。例如，身患罕见疾病的病人、囚犯和前任美国参议员可能非常难找到。研究人员在设计已知群体研究时，需牢记这些限制条件。

所有校标效度研究都有两个重要特征。第一，假设必须有可靠的理论依据。要同时检验理论效度和量表效度是非常困难的。当出现问题时，这种情况常常发生，很难判断问题是源于理论还是源于量表。第二，为了更好地检验量表效度，必须要有好的测量标准。不能找到与标准的预期关系，也可能反映了标准无效，而不是量表无效。在得出有关量表效度的结论之前，必须对标准的效度有足够把握。

聚合和区别效度(Convergent and Discriminant Validity)

聚合效度指同一个概念的不同测量之间存在很强的相关性。区别效度指不同概念的测量之间仅存在适度的相关。这两类效度常常放在一起研究，假设概念内部的关

系和概念之间关系的相对大小。也就是说，一个量表与同一概念的其他测量的相关程度将高于它与其他概念的测量的相关程度。基本的理念是一个概念与自身的相关程度比它与其他概念的相关程度高。

聚合效度能通过比较量表的分值和同一概念的另一个测量来显示。我们预期这两个测量的相关程度很高。理想的情况是，它们的相关性只受其信度水平的影响。也就是说，如果一个概念的两个有效测量都是绝对可信的，那么它们也应该几乎是完全相关的。但是由于总是存在一些误差使得测量的信度降低，所以观察到的同一概念的不同测量之间的相关性在一定程度上也会降低。

评估聚合效度的能力取决于是否存在备用测量。有些概念存在这样的备用测量，但另外一些的概念可能就不存在这样的备用测量。在某些领域，可能已有成熟的量表，可用来当作评估新量表的标准。尽管已有一个差不多能用的量表，但是我们要设计一个新量表，是因为我们要把它用于某一特殊受访群体或某种特别目的。此时，量表的效度可利用一个更一般的标准来检验。

坎贝尔(Campbell)和菲斯克(Fiske)于1959年提出多元特征和多重方法矩阵(Multitrait-Multimethod Matrix，以下简称MTMM)来同时研究聚合效度和区别效度。使用这种方法至少需要测量两个概念，且每个概念至少要用两种不同的测量方法。因为不可能总用多种方法测量同一

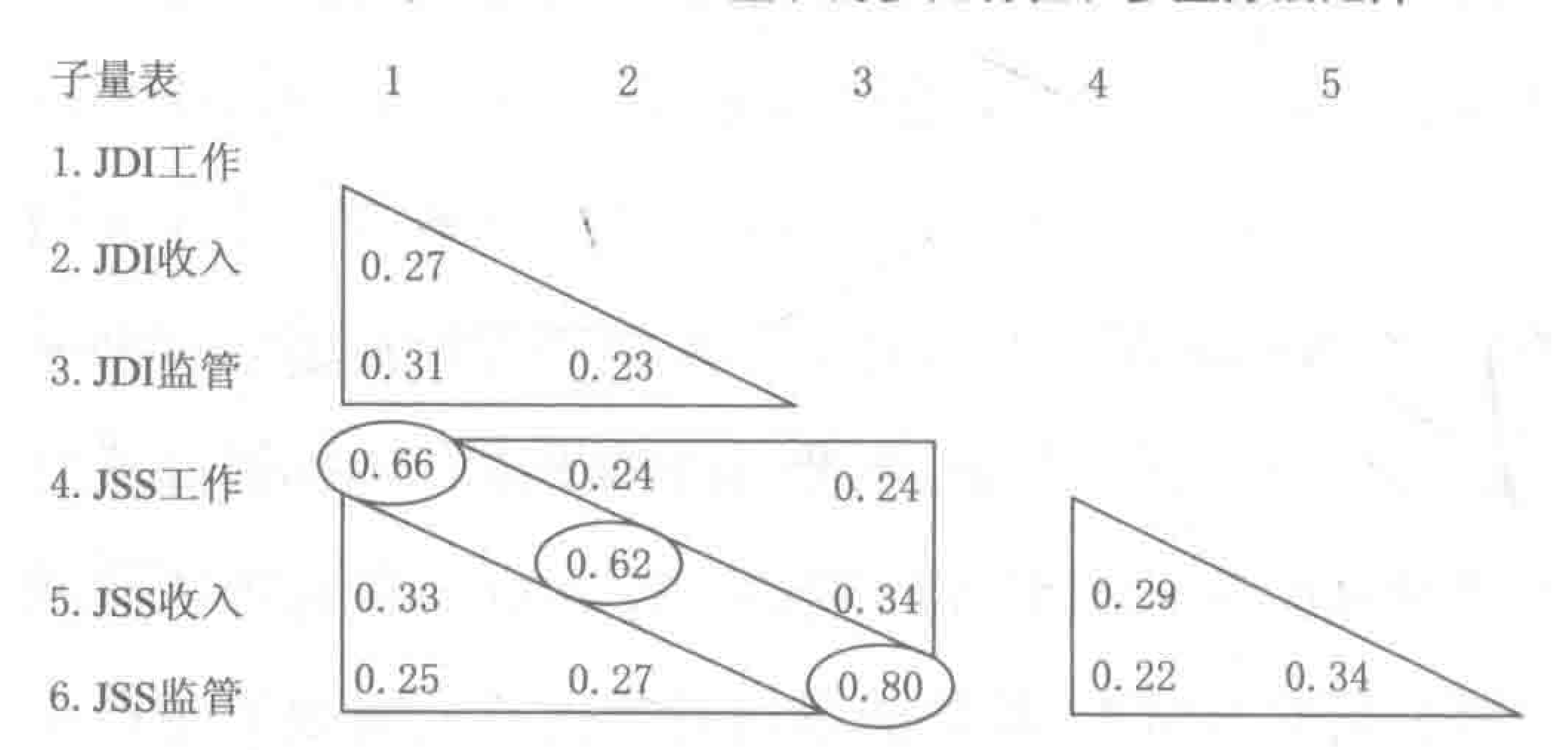
概念，所以第二个要求大大限制了这种方法的用武之地。

这里举例说明 MTMM 这种方法。（详细的讲解及例子参见 Campbell and Fiske，1959）。这个例子来源于一个效度研究，这个研究是比较职员工作满意度的新测量和一个现有的标准。工作满意度调查（the Job Satisfaction Survey，以下简称 JSS；Spector，1985）是一个有 9 个子量表的测量工具，用来评价人们对工作的不同方面的满意度和态度。我们设计这个量表，特别用于像医院、精神病院或社会服务机构等人类服务组织的工作人员。在有关工作满意度的研究领域中，被普遍采用且效度高的量表是工作描述性指标（the Job Descriptive Index，以下简称 JDI；Smith et al.，1969）。JSS 中的 5 个子量表也包含在 JDI 中。这两个量表都应用于一个包括 102 个工作人员的样本（详细介绍参见 Sepctor，1985）。考虑到每个量表作为一种独立的方法，我们可以获得同一受访群体填写两个量表的数据，这样的数据的可获得性允许进行 MTMM 分析。

表 6.1 是个包括 3 个子量表的 MTMM 矩阵。这 3 个子量表分别为对工作任务、工资和上级监管的满意度。这个矩阵包括所有 6 个变量（3 个 JDI 量表和 3 个 JSS 量表）的交互相关。矩阵分析包括把这些相关分为三类。第一类是在每个测量工具内各子量表之间的相关。这些相关是异质同方相关（heterotrait-monomethod correlations）。第二类是各测量工具之间各子量表之间的相关。这些相关

是异质异方相关（heterotrait-heteromethod correlations）。最后一类是测量同一特征但利用不同测量工具的不同子量表之间的相关。这些相关是聚合效度。

表 6.1　三个 JSS 和 JDI 子量表的多元特征和多重方法矩阵



子量表	1	2	3	4	5
1. JDI工作					
2. JDI收入	0. 27				
3. JDI监管	0. 31	0. 23			
4. JSS工作	0. 66	0. 24	0. 24		
5. JSS收入	0. 33	0. 62	0. 34	0. 29	
6. JSS监管	0. 25	0. 27	0. 80	0. 22	0. 34

这些不同类别的相关都呈现在表 6.1 中。在最上面和最右边的三角形内显示的值是异质同方值。最上面的三角形包括 JDI 的子量表之间交互相关。最右边的三角形包括 JSS 的子量表之间的交互相关。位于左下的是两个由画圈的对角线值分开的三角形。这两个三角形内显示的是异质异方相关。这些相关是两个测量工具的不同子量表之间的相关。最后，画圈的值是聚合效度，这些值组成了效度对角线。

聚合效度可通过比较对角线效度值和矩阵中的其他值来确定。这些值应该统计上显著，且值相对较大。对于每个子量表，效度值应该比其同行或同列的其他任何值大。也就是说，同一概念的两个测量工具之间应该比它们与任何其他概念之间的相关更强。

区别效度可通过观察三角区域的相关值比画圈的聚合效度值更小来确定。每个三角形包括待比的每对子量表的相应比较。例如,每个三角形表示工资和监管这两个子量表之间的相关。这些相关(在每个三角形内)的排序应该在不同的三角形中都是相同的。

从表6.1可以看出,JSS的聚合效度和区别效度都很好。对角线效度值都相对较大(从0.62到0.80),且是整个矩阵中的最大值。每个三角形内的对应相关值都很类似且相对较小,从0.22到0.34。这些相对小的值表明不同子量表测量的是不同的概念。

在所有量表的效度都很清楚的情形下,坎贝尔和菲斯克(Campbell and Fiske, 1959)的评估矩阵方法很好用。在其他情形中,任何偏离理想情形的离差都可能使得这种方法不好用。而且,相当多的统计方法可用来评估MTMM矩阵(参见Schmitt and Stults, 1986)。现在最常用的方法也许是结构方程模型(structural equation modeling)(Long, 1983; Widaman, 1985)。这个方法能非常实用,但也有其严重局限性(Brannick and Sepctor, 1990; Marsh, 1989)。最近兴起的直积模型(direct product models)虽然现在还很少在研究文献中(如Bagozzi and Yi, 1990; Wothke and Browne, 1990)看到,但也很有发展势头。为了更好地设计量表,检验矩阵自身能明确地看出在某些子量表中而不是其他子量表中的潜在效度问题。

第2节 | 因子分析在量表效度验证中的应用

因子分析对于检验单维度和多维度量表的效度都能非常管用。对于单维度量表，因子分析能用来探究所选题项组内的可能维度。对于多维度量表，因子分析能用来证实题项从经验上组成了预构建的子量表。

两类基本的因子分析能用于量表的设计。探索性因子分析(exploratory factor analysis)用于确定一组题项中可能存在的不同成分数。验证性因子分析(confirmatory factor analysis，以下也简称 CFA)允许检验一个假设的结构。当我们检验子量表的题项，看它们是否能支持预验证的子量表结构时，我们就可用验证性因子分析。

因子分析的详细讨论远超出本书的讨论范围。不过许多书都介绍了这两类因子分析方法，包括金和米勒(Kim and Mueller，1978a，1978b)的两本书介绍了探索性因子分析，朗(Long，1983)的书中介绍了验证性因子分析。不过，在讨论这些方法在量表设计中的使用时，需要记住这些方

法的一些基本特点。

因子分析的基本思想是把很多题项归纳进少数几个称之为因子(factors)的潜在题项组。例如,一个包括50个题项的量表可归纳进5个因子,每个因子包括10个题项。这些因子可能为不同概念的指标,或者它们是单个相对异质的概念的不同方面的指标。当单独看因子分析的结果时,我们很难确定哪种情形会出现。

因子分析通过分析题项之间的协变(或相关)模式来得到因子。那些与组内各题项的相关比它们与其他组的题项的相关强的题项将组成因子。这有点类似于聚合效度和区别效度。我们假设相关程度相对高的题项反映相同的概念(聚合效度),相关强度相对低的题项反映不同的概念(区别效度)。如果所有的题项之间的相关程度都很强,且大小差不多,那就只会生成单个因子。这表明这个量表测量的是单个概念。

要记住的一个因子分析特征是结果是输入题项的函数。因子能解释的方差的相对强度或比例是所选题项的特征和数量的函数。对于一个多维量表,那些题项多的子量表更可能生成能解释更多方差的更强的因子。题项很少的子量表可能生成的因子解释力非常弱。劣势题项(poor items)和反应偏差(response biases)能严重破坏一个因子方案。

探索性因子分析

探索性因子分析是研究(假设的)单维或多维量表的维度的一个非常好的方法。因为这里分析的目的常常是探索量表的维度,所以虽然还有其他模型可用,但主成分(principal components)分析看起来将是一个合理可用的因子分析模型。

因子分析必须解决两个主要问题:(1)最优代表这些题项的因子数;(2)因子的解释。虽然因子分析是个数学方法,但是这两个问题的答案却落入主观判断和统计决策原则(statistical decision rules)的范围。

分析本身包括几个重复的步骤。首先,我们得到的是主成分,每个分析的题项得到一个主成分或因子。每个初始因子都对应一个特征值(eigenvalue),这个特征值表示被每个因子解释的方差的相对比例。每个题项将对应一个特征值,特征值的总和将等于题项的总数。如果题项之间不相关,那特征值将仅仅反映原始题项的方差,每个特征值都等于1.0。这表明这些题项不能聚合成因子。但当题项之间的相关越来越强时,那它们生成的因子包含越来越多的题项方差,且初始特征值将大于1.0。如果所有题项之间完全相关,那将生成单个因子,其特征值就等于题项的数量,其他特征值为0。如果所有题项生成几个因子,那

每个因子都对应于一个大于 1 的特征值，这表明一个因子能比一个题项解释更多的方差。

一旦确定存在多少个因子，我们将使用正交旋转法（an orthogonal rotation procedure）旋转这些因子。旋转法是设计用来生成基于不同数学标准的题项组（参见 Kim and Mueller, 1978a）。虽然有好几种选择存在，但所有这些选择都最后生成一个（或多个）负载矩阵（a loading matrix or matrices），这些负载矩阵显示每个题项与每个因子之间的相关程度。一个负载矩阵包括统计量（因子负载量，factor loadings），即每个原始变量与每个因子之间的相关。对应于每个因子（矩阵中以列来表示），每个变量（矩阵中以行来表示）有一个负载量。最理想的情形是每个变量仅对应于一个因子，且对应的负载量非常大。当对于一个因子，一个变量有很高的负载量，我们就可以说这个变量“负载”到这个因子上。一个题项负载到任何一个因子的最小值要求至少为 0.30 到 0.35 左右。

因子分析方法的困境在于需要主观判断来确定因子数和因子的解释。有几个步骤来确定因子数（参见 Kim and Mueller, 1978a）。这些步骤包括了反映每个因子解释的方差量的特征值。

许多研究人员使用的策略是多次旋转不同数量的因子，然后依据解释的意义来确定因子数。研究人员可能利用因子分析来分析 30 个题项，然后根据特征值来决定旋转

3、4、5 个因子。在每种情形下，每个因子的题项都要仔细审查，确定是否能构成一个概念上有意义的因子。有可能发现，仅仅三个因子的方案能得到概念上有意义的因子。

在这点上，结果的解释变成一个非常主观且概念化的过程。本来认为是单维度的量表很可能看上去有多个子量表。在统计意义上来看，结果显示基于题项之间的相对关系强度，题项可聚集成多个组。这种情形能发生的原因是量表测量不同的概念，或者量表测量的概念是异质的，包括多个成分。是把这个量表作为单维度用还是作为多维度来用永非易事。

下面以焦虑这个概念为例来说明这点。焦虑包括认知(比如感觉害怕)和身体(比如手心出汗)两方面。一些焦虑量表包含了与这两方面都有关的题项。是该把这两方面分开，还是它们是同一潜在概念的两个指标呢？这个问题的答案取决于研究目的。如果研究目的是调查焦虑对人们完成任务的表现的影响，那么把这两方面分开可能没什么用。但当研究人员研究焦虑的反应本身时，分开这两方面来研究可能就很有意义。

从量表设计的角度来看，因子表示不同概念的结论必须基于效度验证。因子分析能表明因子可能存在，但进一步的效度证据必须收集。在最后的分析中，如果每个因子都有相同的相关性，那这些因子就不能基于校表效度来区分。当因子不能基于它们与其他概念的关系来区分时，很

难说它们表示不同的概念。在这种情形下，简约法则（the law of parsimony）表明应该得出一个更简单的单维度结论。

对于多维量表，研究人员应该从经验上检验因子是否构成原始子量表。为了完成这个任务，研究人员应该选择子量表数作为因子数来旋转。这个方案将显示数据拟合预期量表结构的好坏程度。然而，这需要谨慎处理。因子分析结果很可能对包含在内的所有题项非常敏感。增加或减少单个题项可能严重影响结果。未能找到能完全拟合子量表的因子结构是很正常的。研究人员必须考虑因子结构偏离的程度，看看这是否意味着子量表存在严重问题。

另外，需要注意无意义的因子分析结果。得到一个每个因子只对应于很少题项（例如 1 或 2 个题项）的因子结构一般是没有多大用处的。对于有很少题项的量表，因子分析不可能有多大用。

工作满意度调查（the Job Satisfaction Survey，简称 JSS；Spector，1985）列举了一个例子来说明如何把因子分析应用于多子量表这样的测量工具。这个量表是设计用来测量工作满意度的 9 个方面。根据题项分析的结果，每个子量表选择 4 个题项。我们利用探索性因子分析来确定是否支持 9 个子量表。结果发现只有 8 个特征值大于 1.0。而且，当 9 个因子旋转时，很明显某些子量表分解了。也就是说，每个因子包括的 4 个题项不是来自单个子量表。

检查特征值可看出这个数据至多只代表 8 个因子，并

且对它们进行旋转。因子结构支持 8 个子量表中的 6 个。因子 3—8 只包括 4 个题项，而且对于这 6 个因子中的每一个，这 4 个题项只来自一个子量表。因子 1 和 2 的结果更复杂。这两个因子的负载矩阵及负载在这两个因子上的题项如表 6.2 所示。表 6.2 的第 1 列显示的是题项，第 2 列显示的是每个题项对应的子量表名称。负载在头两个因子的题项仅来自 9 个子量表中的 3 个。这 3 个子量表分别为工资的满意度(PAY)、监管的满意度(SUP)和奖金的一般可获得性(REW)。表 6.2 的第 3 列和第 4 列显示的分别为第 1 和第 2 个因子的负载。反面措辞的题项会出现负的负载值，因为这些题项的分值没有反转。

表 6.2　工作满意度调查(JSS)的收入和监督因子负载值

题　　项	子量表	因子 1	因子 2
我感觉我的工作得到相应的报酬。	PAY	0.07	0.77
我的主管非常胜任他/她的工作。	SUP	0.80	0.05
当我很好地完成工作时，我得到相应的赏识。	REW	0.46	0.25
工资涨得太少且很少涨。	PAY	0.03	−0.54
我的主管待我不公。	SUP	−0.73	−0.06
我觉得我的工作没得到认同。	REW	−0.33	−0.27
当我考虑到公司给我的收入时，我觉得我的工作没得到重视。	PAY	−0.08	−0.73
我的主管很少体谅下属的感受。	SUP	−0.76	−0.08
这里的工作人员很少有奖金。	REW	−0.11	−0.39
我很满意涨薪水的机会。	PAY	0.08	0.56
我喜欢我的主管。	SUP	0.77	−0.01
我感觉我的努力没有得到应有的回报。	REW	−0.17	−0.48

注：SUP = 监督；REW = 奖金；PAY = 收入。

从表6.2中能看出，4个有关监管的题项仅仅负载在因子1上，而4个有关工资的题项仅仅负载在因子2上。而有关奖金的题项则分开。最后两个题项只负载在因子2上。第1个题项仅负载在因子1上。第2个有关奖金的题项则更多地负载在因子1上，但负载在因子1和2上的差异不大(−0.33和−0.27)。注意，这个题项因为使用了“不”来改变措辞方向，所以违反了好的题项的五个原则之一。

这些结果表明JSS包括8个维度，而不是9个维度。一个子量表的题项并入到其他两个子量表中。鉴于这些结论，研究人员可能会倾向于删除奖金的子量表，把其中的题项放到工资和监管的量表中。有趣的是，把这些题项添加到其他两个子量表中并不能明显地提升它们的系数α，即使题项从4个增加到6个。另一种可能性是删除子量表，假设它仅仅测量两个其他子量表的重叠部分(也就是说，这个子量表是多余的、没有必要的)。现在这个子量表还留在测量工具中。如果将来的研究仍未能显示它的用武之地，那将删除这个子量表。

验证性因子分析

验证性因子分析(CFA)允许统计检验一个假设的因子结构。像JSS这样的量表，研究人员可能会假设一个结

构，在这个结构中每个题项都负载到它的子量表上的。CFA 将用来表明数据拟合这个假设结构的程度。

在探索性因子分析中，最好的拟合因子结构是拟合数据。而在验证性因子分析中，结构是事先假设好的，后面用数据来拟合这个假设的结构。在验证性因子分析中，因子负载量的估计如同探索性因子分析中那样。而且，CFA 还得出一个数据拟合程度的指标。

现在做验证性因子分析最好的方法是利用现成的协方差建模程序（covariance structure modeling programs）。两个最常用且普遍可获得的程序是 LISREL（Jöreskog and Sörbom，1984）和 EQS（Bentler，1985）。下面有关 CFA 的评论将仅限于协方差结构建模方法。其详细介绍能在许多文献中找到，例如朗的论文（Long，1983）。

为了做验证性因子分析，因子数、每个题项负载到的那些因子、是否这些因子存在相关都必须提前设定。每个题项负载到每个因子上的负载值呈现在一个负载矩阵中，就像在探索性因子分析中那样。在这个负载矩阵中，行表示各题项，列表示各因子。矩阵中的每个元素要么设为 0，这意味着这个题项没有负载到这个因子上；要么不限定，因而它的负载值能通过分析来估计。为了做这个分析，不限定每个对应于假设题项的因子所处的元素，限定所有其他元素都为 0。

也有一个表示因子之间相关的矩阵。这些相关也能

限定为0,这样这个分析就成了一个正交或不相关的因子分析。另一种选择是不限定这些相关值,这样能让程序计算各因子之间的相关。

限不限定元素后面的逻辑在于,研究人员能限定那些总体中假设为0的相关(或因子负载量),而不限定那些(绝对值)预期大于0的相关(或因子负载量)。这种"模型设定"可能会出现两种错误:限定的负载值可能实际上不等于0,或未限定的负载值可能等于0。如果因子结构设定正确,即因子数及限定和不限定参数的模式都正确,那么数据应该拟合得很好。当然,由于抽样误差,特别是小样本,数据可能无法与设定的模型完全拟合。

验证性因子分析将得到因子负载值和因子之间的相关系数的估计值。分析也给出每个不限定参数的显著性指标(在LISREL中是T值)。所有不限定的值都预期是相对较大且显著的。分析也将给出数据总体拟合模型的几个指标。

总体拟合和单个参数都应该检查得出有关子量表结构是否合适的结论。有可能某些子量表很好,而另一些子量表则不然。JSS的探索性因子分析就出现了这种情况。也有可能总体的结构得到一个很好的拟合指标,虽然许多负载值很小且不显著。实际上,错误地不限定某个参数对于总体拟合的影响要小于错误地限定某个不该限定的参数。这是因为错误地不限定的那个参数也能被估计到其

正确值(0)。但是,当参数值被错误地限定为0时,那么可能限定值与真实值相差甚远。

拟合得好的验证性因子分析表明子量表结构可能解释数据,但这并不意味着这个结构确实解释了数据。换句话说,支持一个测量工具中的子量表并不是它们反映待测概念的强有力的证据。得出这样的结论还必须基于本章讨论其他类型的证据。

第 3 节 | WLCS 的效度

WLCS 的效度验证说明了量表设计人员一般采用的方法。主要的效度验证依据通过计算 WLCS 的分值和与之预期相关的几个标准之间的相关来提供。因子分析用来探究这个量表的维度。

校标效度数据由采用它们的 6 项研究提供(Spector, 1988 中总结了这些研究)。这些研究提供了 WLCS 和 10 项标准之间的相关系数。除了一种情况外,每个标准起码用于两个以上的研究。表 6.3 给出了校标效度结果的总结。表中列出了每个变量的样本数量、所有样本的总样本量和相关均值。

WLCS 预期相关的第一个标准是一般控制点(general local of control)。这是因为工作控制点是一般控制点理论的主要运用理论。从理论上来说,在一般控制点上表现为内控型(或外控型)的人应该倾向于在个体领域上的得分同向。因为假定这两类控制点稍有不同,这两个测量之间的相关应该不是很大。也就是说,这应该不影响量表的信

表 6.3 工作控制点(WLCS)和标准之间的相关总结

标　　准	样本数	受访者数	相关均值
一般控制点	3	800	0.54
工作满意度	5	968	−0.54
组织归属感	3	222	−0.24
意图辞职	5	667	0.23
工作自主性	2	579	−0.13
决策影响力	3	220	−0.37
角色压力	1	287	0.32
主管体贴	3	182	−0.31
主管初始结构	2	133	−0.33

度。从表 6.3 中可看出,这 3 个样本之间的平均相关为 0.54。因此,WLCS 与一般控制点相关,但这相关不足以质疑其区别效度。

选择其他的标准是因为这些标准预期与工作控制点相关。基于有关工作环境的一般控制点研究的综述(O'Brien, 1983; Spector, 1982),选择那些预期与工作控制点相关的变量。具体地说,内控型人预期在工作中有更多的自主性和影响力,工作态度更好,角色压力更小,经过更周全的考虑之后才向上级汇报。如表中所示,这些变量之间相关的方向和预期的一样。总体来说,这验证了 WLCS 的效度。

然而,一个棘手的结论是工作控制点与自主性之间的相关很弱。因为自主性是与控制有关的变量,研究人员应该预期内控型人可能比外控型人有更多的自主性。研究

人员也许会猜测这个弱相关的原因(例如,也许工作自主性是如此明确清晰,以至于每个人把它都看成是相同的)。应该做更多的研究来确定,缺乏更好支持这个假设的证据是否对于量表的效度或假设构成问题。

我们利用来自原始样本的数据,对 WLCS 进行探索性因子分析。这里总结了最后 16 个题项的一些结果。表 6.4 包括特征值,这些特征值反映了每个因子解释方差的相对比例。表的最后一列是实际比例。从表中能看出,单个因子甚至不能解释这些题项三分之一的方差。这表明需要更多的因子。

表 6.4 工作控制点量表(WLCS)的 16 题项因子分析的特征值

因子数	特征值	方差比例
1	5.08	31.8
2	2.27	14.2
3	1.21	7.6
4	0.97	6.0
5	0.86	5.4
6	0.79	5.0
7	0.73	4.6
8	0.69	4.3
9	0.63	4.0
10	0.54	3.4
11	0.46	2.9
12	0.44	2.7
13	0.42	2.6
14	0.35	2.2
15	0.30	1.9
16	0.25	1.6

当两个因子旋转时，结果变得相当有趣。表 6.5 包括每个题项在两个因子上的负载值，其中也包括题项的措辞方向。斜体值表示每个题项负载最好的因子。这个因子结构非常清晰，因为外向型的题项组成第一个因子，而内向型的题项组成第二个因子。

表 6.5 工作控制点量表(WLCS)的两个因子方案的因子负载值

题项数	措辞方向	因子 1	因子 2
1	内控型	0.02	*0.67*
2	内控型	−0.04	*0.59*
3	内控型	−0.16	*0.52*
4	内控型	−0.02	*0.52*
5	外控型	*0.70*	−0.05
6	外控型	*0.73*	−0.13
7	内控型	−0.02	*0.67*
8	外控型	*0.74*	0.03
9	外控型	*0.78*	−0.15
10	外控型	*0.72*	−0.30
11	外控型	*0.70*	−0.08
12	内控型	−0.14	*0.53*
13	外控型	*0.77*	−0.14
14	外控型	*0.67*	−0.16
15	内控型	−0.23	*0.65*
16	内控型	−0.27	*0.52*

内向型子量表分值通过加总所有内向型措辞的题项得到，而外向型子量表分值通过加总所有外向型措辞的题项得到。在原始样本中，它们之间的相关为−0.32。这个相关和因子分析结果强烈表明工作控制点由两个稍稍相互独立的部分——内控型和外控型组成。基于因子分析

结果好几个控制点的研究人员都得出关于一般控制点的相同结论。

但是，如前面讨论的那样，因子分析结果不足以得出因子代表独立概念的结论。需要附加的数据来证实这些不同的部分。我们为 WLCS 收集了这些附加的数据。单独的内控型和外控型子量表与两个单独样本的标准相关。总之，计算了 32 对相关值。这两个子量表之间的相关的相应大小利用 t 检验进行统计比较。32 个相关中只有 1 个显示了两个量表的校标效度系数不同。如果多重检验已经调整了实验误差，那么发现显著差异的敏感性应该可能更小。这个显著的相关非常有可能是由抽样误差引起的。

当两个子量表都与同一标准相关时，结果的收敛表明这些数据只代表一个概念。然而，因子分析结果不支持这个结论。还有一个证据也许能帮助解决这个分歧。

将内控型和外控型子量表的分值都画在一个散点图上。这个图显示了两个子量表的取值范围。对于一个有 8 个题项、每个题项有 6 个答案选项的量表，分值的取值范围从 8 到 48。内控型分值的取值范围从 26 到 48，很少分值低于 30。外控型分值的取值范围从 8 到 37，很少值高于 35。在这个例子中，很少受访者的外控型分值很高。取值范围的限定可能是减弱子量表之间相关的原因。对于一个外控型受访者更多的样本，预期结果可能不同。

WLCS 的效度数据当即是非常有价值的，但它们仅仅

是开端。我们可能需要通过把量表应用到包括更多外向型人的总体中得到更强的效度证据。附加的校标效度研究也应该进行。最后,内控型和外控型题项反映不同概念的可能性应该做进一步的研究,也许通过用其他标准继续研究不同的效度。

第4节 | 效度策略

检验效度需要一个尽可能多地收集不同种类的证据的策略。一个有效的策略是在量表设计初期就开始研究效度了。最初的题项备选库的首次应用可能包括能用于效度验证的附加标准变量的测量。如果从最初题项备选库得到一个尝试性量表版本,有关效度验证的数据将可获得的。然后,研究人员将检验看是否新量表与假设的附加标准变量相关。如果样本量足够大,因子分析也能做。

假设量表看起来很有价值,附加效度验证研究就要继续进行。这些可能包括重复早期效度检验和一些新检验。那应该尽可能多地做不同类型的检验。研究人员不应该完全依赖自评问卷(self-report questionnaires)来验证一个量表的效度。仅仅发现研究的量表与应用于同一批人的其他量表之间的相关本身并不是一个很强的证据。如果与聚合效度的相关是基于非常不同的操作化(operationalization),那么证据将变得更强有力些。

当有关效度的证据开始逐渐增加时,量表看起来像预

期那样可行将变得更明显。在这点上,研究人员能得出结论说,检验说明了概念效度。然而,要记住,概念效度能被支持,但永远不能被证明。与任何科学理论一样,总是有可能以后的工作将重新解释之前的研究结果,然后发现先前的解释不正确。这样,量表评估的概念本质将被赋予一个新的解释。然而,直到那天到来,之前量表的概念仍然假定是有效的。

第7章

信度和标准

最后两个要讨论的问题是测量工具的信度构建和标准编辑。虽然这两者在最后讨论，但与这两者有关的数据在量表应用过程中都收集了。

信度在介绍题项分析的处理过程中也涉及一些。题项分析包括了系数 α 的计算和内部一致性信度的建立。进一步的工作应该确认测量工具的内部一致性在其他样本中成立，并建立检验—再检验信度。

确定量表的总体均值和标准差的估计值也是我们感兴趣的问题。这个信息在确定分值的意义时非常有用。整个标准检验的方法是，基于确定相对于总体的其他成员，个人所处的位置。根据大部分总体得分的位置来确定个人的分值是高或是低。

第 1 节 | 信度

内部一致性信度是一个量表的各题项如何反映一个普通潜在概念的指标。虽然系数 α 不是唯一可用来评估内部一致性的统计量,但它却是最常用的。即使在开始的题项分析中得到非常高的内部一致性,也最好根据随后的样本重新再作计算。信度估计值在不同类型样本中的可获得性,将拓展量表信度应用到一个更广泛群体的普适性。每一次使用量表时都计算系数 α 是个明智之举。不同样本计算得到的系数 α 值应该变化不大。但是,当一个量表的内在一致性出现问题时,那不同样本计算得来的系数 α 值可能成倍变化。基于这个原因,我们应该每次都检查这个统计量。使用计算机和合适的软件,计算系数 α 是非常简单的。

检验—再检验信度也应该确定。这类信度反映了随时间推移测量的一致性。这是当量表重复用于同一批被访者时得到的与自身的相关程度。内在一致性的量表常常也有好的检验—再检验信度,但随时间推移的一致程度

必须通过经验验证。我们预期稳定概念的测量随时间推移呈现很高的一致性，但当量表是用来测量随时间变化的概念时，则可能出现例外。一个人情绪（例如高兴或悲伤）的测量是用来评估人们在某个特定时刻的感受。由于情绪能变化相当快，除非时间间隔非常短，否者检验—再检验信度可能很低。信度数据应该根据潜在概念的预期一致性来解释。例如，相对于用来测量更短暂的情绪的量表来说，用来测量一个持久个性特征的量表应该预期呈现更高的检验—再检验信度。

许多量表都计算不同时间间隔的检验—再检验信度。时间间隔越长，信度应该预期越低。态度经过1—2周应该还是相当可信的，但经过20年则应该不太可信了。不过研究表明，即使经过数十年，也有许多概念的信度依然非常高。

检验—再检验信度很容易计算。若一个量表用于同一群受访者两次，姓名或某些独特的个体识别码必须匹配每个受访者的量表。这样分值能在两次量表的应用之间匹配。然后计算这两次量表应用得到的分值之间的相关系数。

第2节 | WLCS的信度

WLCS在初始样本中的内在一致性信度系数为0.85，这是在可接受范围内。内在一致性的可复制性是必要的，这样才能确保这些题项可继续聚合到最后只剩16个题项（而不是所有的49个题项）的其他样本中。而且，因为初始样本由学生构成，而量表应用的目标群体是工作人员，所以有必要证明这个量表在工作人员的样本中的内在一致性。

量表的系数α根据5个由工作人员组成的附加样本重新做了计算（参见Spector，1988）。这些样本代表的工作人员范围很宽，包括管理和非管理阶层，国有和私有部门。这些样本的异质性使得内在一致性能更适应于不同工作人群。这5个样本计算得到的系数α从0.75到0.85，均值为0.82。内在一致性看起来满足不同样本的需要，且是一致的。当量表用于其他样本和其他情形时，内在一致性需要继续监测。

WLCS仅有有限的检验—再检验信度数据。一个包

括31个大学生的小样本填了两次 WLCS 量表,期间间隔了6个月左右。他们在大学的最后一个学期第一次填了 WLCS 量表,然后在毕业后刚开始工作的几个月后第二次填了量表。这个检验—再检验信度系数为0.70。这个样本量非常小,仅仅给出了检验—再检验信度的大致估计。另一个样本量超过100的类似研究现在正在进行。

第3节 | 标准

为了解释分值的意思,了解有关分值在不同人群中的分布将非常有帮助。社会科学中的大部分概念的测量量表是任意的。一个分值的意思只能根据某个参考框架来确定。标准方法是用分值的分布来作为参考框架,这也是大部分社会科学测量的基础。因此,一个人的分值是对照分值的分布得来的。如果个人的分值大于分布的大部分值,那么这个分值被认为是高的;如果分值小于分布的大部分值,那么这个分值被认为是低的。

为了确定总体分布的特征,有必要把量表应用到一个有代表性的大样本中。为了推广到所有美国人这个总体,应该对整个总体进行代表性的抽样。令人遗憾的是,很少社会科学家能轻易地获得一般总体的信息。大部分量表是根据有限的总体发展起来并设定标准的。大部分量表都是根据大学生来设定标准的。对于某些概念,大学生群体可以很好地代表一般总体。但对于其他概念,大学生群体的代表性可能就非常差了。遗憾的是,很难事前知道大

学生的代表性怎么样。这就引出了一个问题，即把量表从实际抽样人群推广到其他人群的一般化问题。

为了编写标准，研究人员要尽可能多地收集不同受访者填写量表的数据。当然，信度和效度研究使用的数据也能加到量表的标准中。一系列详细的类似研究应该能为标准的发展提供一个好的起点。

为了编写标准，研究人员要计算所有受访者的描述性统计量。均值和标准差是主要感兴趣的统计量。分布的形状也应该检验，包括偏斜度和变量取值范围的可能限定等。

如果不同样本的标准要合并，那么我们应该注意各样本的分布要相似。如果样本本身不同，这更需要注意。时间间隔较短的来自同学校同专业的几个大学生样本数据能适当地合并在一起。大学生和非大学样本，男女比例不同的样本，或不同种族的样本要合并的话就要特别谨慎。

这还涉及亚群体标准的问题。某些亚群体(如男性相对于女性，白人相对于非白人)可能在某些量表上表现不同。这些差异常常受到关注，而且许多量表设计者为不同亚群体编写不同标准。性别和种族是两个非常明显区分总体的变量。也可以根据感兴趣概念的本质来推荐其他变量。研究人员应该非常谨慎地研究那些看起来不同的亚群体。当然，要研究所有可能的亚群体是不可能的。

一个严谨的量表设计者应该根据尽可能多样化的群体

来编写标准。克劳尼和马洛(Crowne & Marlowe，1964)为几个不同的样本(包括大学生、求职人员和监狱犯人)提供了 SD 量表的标准。他们继续把这些群体分为男女组。不同群体的标准数据的可获得性增加了有意义的比较的可能性。

即使亚群体存在，研究人员可能仍希望为总体确定一个全面标准。为了完成这个目标，研究人员必须注意，样本要确实代表感兴趣的总体。对于大部分亚群体，样本的构成应该和总体的一样。例如，对于一般总体，男女比例应该近似相等，或许女性比例略高一些。如果总体是高层企业总裁，那有代表性的样本应该大部分都是男性。当然，如果企业总裁存在男女差异，那么男女数据应该用来制定不同的标准。

第 4 节 | WLCS 的标准

WLCS 的标准是根据包括 1 360 个受访者的 7 个样本制定的。其中 6 个样本是前面讨论的信度和效度的数据。剩下的那个样本由 195 个大学生组成。测量工具的均值为 38.1，标准差为 9.4。注意这个测量工具的可能取值范围是 16 到 96，中位值为 56。这使得这个测量工具的均值位于量表的中间值的 1.9 个标准差以下。也就是说，来自抽样总体的受访者选择量表的极端项非常少。

WLCS 遇到的取值范围受限的问题可能是由相当小范围的抽样总体引起的。填写这个量表的大部分受访者要么是大学生，要么是白领。一个包含更多类型受访者的样本将可能出现更多的分值较高的外控型的人。

WLCS 的标准没有根据亚群体来区分。大学生和全职人员可获得的样本之间量表得分没有差异。下一步的努力方向应该确定其他亚群体的得分是否存在差异。特别让人感兴趣的是，找到一个受访者得分趋向外控型的总体。从已有样本得到的取值范围限制不利于量表的假设检验。如果样本中没有外控型的受访者，那研究人员不能比较内控型和外控型受访者。



结 语

量表的设计是个不间断的持续过程。在很大程度上，这是因为大部分概念是嵌入在理论框架中的理论抽象。一个概念的有效性受到它所用理论的有效性的限制。就像理论一样，我们永远不能证明一个量表确实是测量感兴趣的概念。但是，我们能证明一个量表能与它的理论框架按照一致的方式来表现。这使得量表及其概念对社会科学研究人员和实际工作人员来说都非常有用。我们必须认识到，未来的研究可能反驳量表的理论依据和概念效度。

一个重要的问题是决定一个新量表何时可以启用。这没有严格的硬性规定，而且研究人员常常迫于需要，在量表构建初期就开始用了。无论如何，在量表用于任何用途前，应该做题项分析并且得到一个可接受的系数 α。而且，强烈建议在量表使用前，起码收集一些能用于效度验证的数据。信度并不能保证效度。最理想的是，收集几类效度验证的数据。数据的早期使用应该有两个目的，其一

是效度研究。

本书涵盖了构建一个评分加总量表的所有必要步骤。如前面所提到的，不推荐没有经验的人员尝试仅仅依据本书所列的步骤生搬硬套地设计量表。虽然本书已包括所有步骤，但许多步骤的讨论比较简单。这里的一些讨论，包括检验理论、信度、效度和因子分析等需进一步参考其他文献。(前面已列出一些参考文献)对于初次构建量表的人员，建议最好咨询一个熟练掌握量表构建的人。对这些讨论和帮助有更为细致的理解，将有助于更成功地构建量表。

注释

[1] 心理物理学是研究物理刺激和对这些刺激的感知之间关系的学科。在心理物理学中，临界值被定义为处于足够强到能意识到与太弱意识不到之间的边界上的刺激强度。这根据一个人能半数正确报告意识到的刺激强度来操作化。

[2] 我想感谢迈克尔·T.布兰尼克(Michael T.Brannick)讲述这个有关过度阐释因子分析的危害性的故事。

参考文献

ALLEN, M. J., and YEN, W. M. (1979) Introduction to Measurement Theory. Monterey, CA: Brooks/Cole.

BAGOZZI, R. P., and YI, Y. (1990) "Assessing method variance in multitrait-multimethod matrices: The case of self-reported affect and perceptions at work." Journal of Applied Psychology, 75: 547-560.

BENTLER, P. M. (1985) Theory and Implementation of EQS, a Structural Equations Program. Los Angeles: BMDP Statistical Software, Inc.

BRANNICK, M. T., and SPECTOR, P. E. (1990) "Estimation problems in the block-diagonal model of the multitrait-multimethod matrix." Applied Psychological Measurement, 14: 325-339.

CAMPBELL, D. T., and FISKE, D. W. (1959) "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin, 56: 81-105.

CARMINES, E. G., and ZELLER, R. A. (1979) Reliability and Validity Assessment. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-017. Beverly Hills, CA: Sage.

CRONBACH, L. J. (1951) "Coefficient alpha and the internal structure of tests." Psychometrika, 16: 297-334.

CROWNE, D. P., and MARLOWE, D. (1964) The Approval Motive. New York: John Wiley.

EBEL, R. L. (1969) "Expected reliability as a function of choices per item." Educational and Psychological Measurement, 29: 565-570.

GUILFORD, J. P. (1954) Psychometric Methods. New York: McGraw-Hill.

JÖRESKOG, K. G., and SÖRBOM, D. (1984) LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods (3rd ed.). Mooresville, IN: Scientific Software.

KIM, J., and MUELLER, C. W. (1978a) Factor Analysis Statistical Methods and Practical Issues. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-014. Beverly Hills, CA: Sage.

KIM, J., and MUELLER, C. W. (1978b) Introduction to Factor Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-013. Beverly Hills, CA: Sage.

LIKERT, R. (1932) "A technique for the measurement of attitudes." Archives of Psychology, 22: No. 140.

LONG, J. S. (1983) Confirmatory Factor Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-033. Beverly Hills, CA: Sage.

MARSH, E. W. (1989) "Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. Applied Psychological Measurement, 13: 335-361.

NEWSTEAD, S. E., and COLLIS, J. M. (1987) "Context and the interpretation of quantifiers of frequency." Ergonomics, 30: 1447-1462.

NICHOLLS, J. G., LICHT, B. G., and PEARL, R. A. (1982) "Some dangers of using personality questionnaires to study personality." Psychological Bulletin, 92: 572-580.

NUNNALLY, J. C. (1978) Psychometric Theory (2nd ed.). New York: McGraw-Hill.

O'BRIEN, G. E. (1983) "Locus of control, work, and retirement," in H. M. Lefcourt (ed.) Research in Locus of Control (vol. 3). New York: Academic Press.

PHARES, E. J. (1976) Locus of Control in Personality. Morristown, NJ: General Learning Press.

RORER, L. G. (1965) "The great response-style myth." Psychological Bulletin, 63: 129-156.

ROTTER, J. B. (1966) "Generalized expectancies for internal versus external control of reinforcement." Psychological Monographs, 80(1), Whole no. 609.

SCHMITT, N., and STULTS, D. M. (1986) "Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10: 1-22.

SMITH, P. C., KENDALL, L. M., and HULIN, C. L. (1969) The Measurement of Satisfaction in Work and Retirement. Chicago: Rand McNally.

SPECTOR, P. E. (1976) "Choosing response categories for summated rating scales." Journal of Applied Psychology, 61: 374-375.
SPECTOR, P. E. (1980) "Ratings of equal and unequal response choice intervals." Journal of Social Psychology, 112: 115-119.
SPECTOR, P. E. (1982) "Behavior in organizations as a function of employee's locus of control. Psychological Bulletin, 91: 482-497.
SPECTOR, P. E. (1985) "Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey." American Journal of Community Psychology, 13: 693-713.
SPECTOR, P. E. (1987) "Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem?" Journal of Applied Psychology, 72: 438-443.
SPECTOR, P. E. (1988) "Development of the Work Locus of Control Scale." Journal of Occupational Psychology, 61: 335-340.
SPSS Inc. (1988) SPSS-X User's Guide (3rd ed.). Chicago: Author.
WIDAMAN, K. F. (1985) "Hierarchically nested covariance structure models for multitrait-multimethod data." Applied Psychological Measurement, 9: 1-26.
WOTHKE, W., and BROWNE, M. W. (1990) "The direct product model for the MTMM matrix parameterized as a second order factor analysis model." Psychometrika, 55: 255-262.

译名对照表	
a loading matrix or matrices	负载矩阵
a theoretical construct	理论概念
acquiescence response set	默认反应定式
agreement	同意
an orthogonal rotation procedure	正交旋转法
attitudes	态度
bias	偏差
classical test theory	经典检验理论
conceptualization	概念化
concurrent validity	同时效度
confirmatory factor analysis(CFA)	验证性因子分析
construct	概念
convergent validity	聚合效度
covariance structure modeling programs	协方差建模程序
criterion-related validity	校标效度
Cronbach's alpha	科隆巴赫的 α
Crowne-Marlowe SD scale	克罗恩—马洛社会期望量表
deductive approach	演绎式方法
dimensional validity	维度效度
dimensionality	维度
direct product models	直积模型
discriminant validity	区别效度
distributional characteristics	分布特征
eigenvalue	特征值
evaluation	评价
exploratory factor analysis	探索性因子分析
external	外控型
factor analysis	因子分析
factor loadings	因子负载量
force choice	强行选择
frequency	频率

general local of control	一般控制点
heterotrait-heteromethod correlations	异质异方相关
heterotrait-monomethod correlations	异质同方相关
homogeneity	同质性
inconsistency	不一致
inductive approach	归纳式方法
intercorrelate	相关
internal	内控型
internal consistency	内在一致性
internal-consistency reliability	内在一致性信度
item pool	题项备选库
Item-remainder coefficients	余项系数
item-whole coefficient	题项—总体系数
job satisfaction	工作满意度
known-groups validity	已知群体效度
Likert-type items	利克特式测量项
mean	均值
measurement sensitivity	测量敏感度
multiple-item scales	多项量表
Multitrait-Multimethod Matrix(MTMM)	多元特征和多重方法矩阵
norms	标准
observed score	观察值
operationalization	操作化
opinions	观点
part-whole coefficient	部分—总体相关系数
personalities	个性
poor items	劣势题项
population	总体
precision	精确性
predictive validity	预测效度
principal components	主成分
principle of parsimony	简约原则
psychometric properties	心理测量

psychophysical	心理物理学
random error	随机误差
reinforcement	强化
reliability	信度
response biases	反应偏差
response sets	反应定式
Rotter's general locus of control scale	罗特的一般控制点量表
scale reliability	量表信度
scope	范围
self-report questionnaires	自评问卷
social desirability	社会期望
Spearman-Brown prophesy formula	斯皮尔曼—布朗预测公式
standard deviation	标准差
statistical decision rules	统计决策原则
statistical power	统计功效
structural equation modeling	结构方程模型
subconstruct	子概念
subject	受访者
subscale	子量表
subjective social class	社会阶层认同
summated rating scale	评分加总量表
systematic influence	系统性影响
test-retest reliability	检验—再检验信度
the item analysis	题项分析
the job descriptive index(JDI)	工作描述性指标
the job satisfaction survey(JSS)	工作满意度调查
the law of parsimony	简约法则
theory-testing	理论检验
true score	真实值
unreliability	不可靠
validation of the scale	量表的效度
validity	效度
work locus of control scale	工作控制点量表

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic回归入门
39. 解释概率模型：Logit、Probit以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库